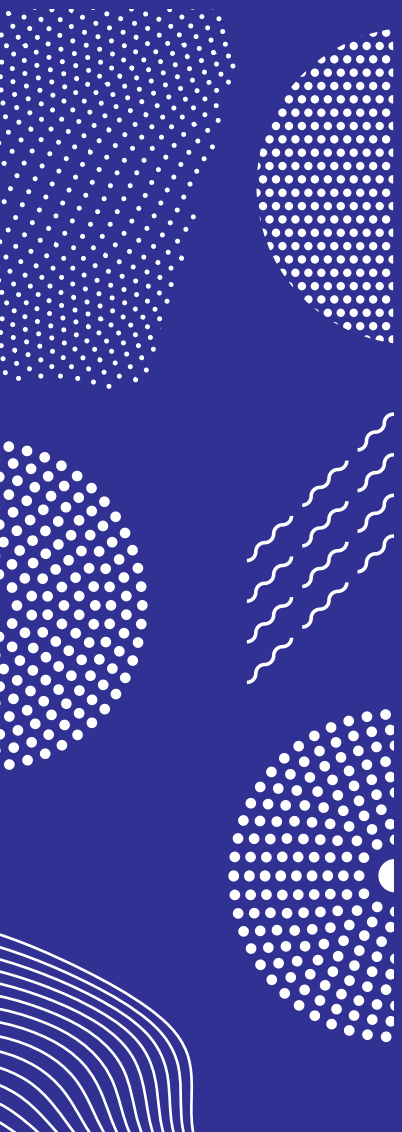
ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

RAPORTTEJA
RAPPORTER
REPORTS
2021:7

APPLICABILITY OF CMIP6 MODELS FOR BUILDING CLIMATE PROJECTIONS FOR NORTHERN EUROPE

KIMMO RUOSTEENOJA



RAPORTTEJA
RAPPORTER
REPORTS

No. 2021:7

Applicability of CMIP6 models for building climate projections for northern Europe

Kimmo Ruosteenoja

Ilmatieteen laitos
Meteorologiska Institutet
Finnish Meteorological Institute

Helsinki 2021

ISBN 978-952-336-141-6 (pdf)

ISSN 0782-6079

DOI: 10.35614/isbn.9789523361416

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Published by	**Finnish Meteorological Institute** (Erik Palménin aukio 1), P.O. Box 503 FIN-00101 Helsinki, Finland	Series title, number and report code of publication Reports 2021:7 Date 16 September 2021
Author(s) Kimmo Ruosteenoja ORCID iD 0000-0002-4370-0782		Name of project HEATCLIM, CHAMPS, FINSCAPES, LEGITIMACY

Title: Applicability of CMIP6 models for building climate projections for northern Europe

Abstract

In this report, we have evaluated the performance of nearly 40 global climate models (GCMs) participating in Phase 6 of the Coupled Model Intercomparison Project (CMIP6). The focus is on the northern European area, but the ability to simulate southern European and global climate is discussed as well.

Model evaluation was started with a technical control; completely unrealistic values in the GCM output files were identified by seeking the absolute minimum and maximum values. In this stage, one GCM was rejected totally, and furthermore individual output files from two other GCMs. In evaluating the remaining GCMs, the primary tool was the Model Climate Performance Index (MCPI) that combines RMS errors calculated for the different climate variables into one index. The index takes into account both the seasonal and spatial variations in climatological means. Here, MCPI was calculated for the period 1981—2010 by comparing GCM output with the ERA-Interim reanalyses. Climate variables explored in the evaluation were the surface air temperature, precipitation, sea level air pressure and incoming solar radiation at the surface.

Besides MCPI, we studied RMS errors in the seasonal course of the spatial means by examining each climate variable separately. Furthermore, the evaluation procedure considered model performance in simulating past trends in the global-mean temperature, the compatibility of future responses to different greenhouse-gas scenarios and the number of available scenario runs. Daily minimum and maximum temperatures were likewise explored in a qualitative sense, but owing to the non-existence of data from multiple GCMs, these variables were not incorporated in the quantitative validation.

Four of the 37 GCMs that had passed the initial technical check were regarded as wholly unusable for scenario calculations: in two GCMs the responses to the different greenhouse gas scenarios were contradictory and in two other GCMs data were missing from one of the four key climate variables. Moreover, to reduce inter-GCM dependencies, no more than two variants of any individual GCM were included; this led to an abandonment of one GCM. The remaining 32 GCMs were divided into three quality classes according to the assessed performance. The users of model data can utilize this grading to select a subset of GCMs to be used in elaborating climate projections for Finland or adjacent areas. Annual-mean temperature and precipitation projections for Finland proved to be nearly identical regardless of whether they were derived from the entire ensemble or by ignoring models that had obtained the lowest scores. Solar radiation projections were somewhat more sensitive.

Publishing unit: Weather and Climate Change Impact Research

Classification (UDC) 551.581.1, 551.58, 551.583,551.588.7, 519.673	Keywords CMIP6 models, SSP scenarios, model evaluation, MCPI index, ERA-Interim reanalyses, northern Europe
ISSN and series title: 0782-6079 Reports	ISBN 978-952-336-141-6
DOI https://doi.org/10.35614/isbn.9789523361416	Language English (abstract also in Finnish and Swedish) Pages 48

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Julkaisija	**Ilmatieteen laitos**	Julkaisun sarja, numero ja raporttikoodi
	(Erik Palménin aukio 1)	Raportteja 2021:7
	PL 503, 00101 Helsinki	Päiväys 16.9.2021
Tekijä(t)		Projektin nimi
Kimmo Ruosteenoja		HEATCLIM, CHAMPS, FINSCAPES, LEGITIMACY
ORCID iD 0000-0002-4370-0782		

Nimeke: Maailmanlaajuisten CMIP6-ilmastomallien soveltuvuus Pohjois-Euroopan ilmastonmuutosskenaarioitten laatimiseen

Tiivistelmä

Raportissa on arvioitu lähes 40 maailmanlaajuisen CMIP6-ilmastomallin (CMIP6 = Coupled Model Intercomparison Project, 6. vaihe) toimivuutta. Arviointi painottuu Pohjois-Euroopan alueelle, mutta lisäksi on tarkasteltu myös mallien kykyä simuloida Etelä-Euroopan ja koko maapallonkin ilmastoa.

Mallien arvioiminen aloitettiin teknillisellä ennakkotarkastuksella; mallien tulostiedostoissa esiintyvät täysin epärealistiset arvot saatiin selville etsimällä kustakin tiedostosta absoluuttisesti pienimmät ja suurimmat arvot. Tässä vaiheessa yksi malli jouduttiin hylkäämään kokonaan, ja sen lisäksi kahden muunkin mallin yksittäiset tulostiedostot osoittautuivat kelvottomiksi. Jäljelle jääneitä malleja arvioitaessa ensisijainen työkalu oli MCPI-indeksi (MCPI = Model Climate Performance Index), joka kokoaa eri ilmastomuuttujille lasketut normitetut RMS-virheet yhdeksi indeksiksi. Indeksi ottaa huomioon sekä ilmastosuureitten vuodenaikaiset että alueelliset vaihtelut. MCPI-indeksi laskettiin jaksolle 1981—2010 vertaamalla mallien tuloksia ERA-Interim-uusanalyyseihin. Arvioinnissa käytetyt ilmastomuuttujat olivat pintalämpötila, sademäärä, merenpinnan tasolle redukoitu ilmanpaine ja pinnalle tuleva auringonsäteily.

MCPI-indeksin ohella tutkittiin erikseen jokaisen ilmastomuuttujan aluekeskiarvojen RMS-virheitä. Lisäksi arvioinnissa otettiin huomioon mallien kyky simuloida maapallon keskilämpötilan tähänastisia muutoksia, eri kasvihuonekaasuskenaarioita vastaavien tulevien muutosten keskinäinen yhteensopivuus sekä käytettävissä olevien skenaarioajojen lukumäärä. Vuorokauden alimpien ja ylimpien lämpötilojen hyvyyttä tutkittiin ainoastaan kvalitatiivisesti, sillä näistä suureista tiedot puuttuivat varsin monesta mallista; kyseisiä muuttujia ei siis otettu mukaan hyvyysindeksejä laskettaessa.

Neljä mallia niistä 37:stä, jotka olivat läpäisseet alustavan teknillisen tarkastuksen, osoittautui eri syistä täysin sopimattomiksi skenaariolaskelmiin. Kahdessa mallissa eri kasvihuonekaasuskenaarioitten tuottamat vasteet olivat ristiriidassa keskenään, ja kahdessa muussa mallissa tiedot jostakin neljästä keskeisestä ilmastomuuttujasta puuttuivat. Lisäksi mallien välisten riippuvuuksien vähentämiseksi mukaan kelpuutettiin kustakin mallista enintään kaksi versiota; tämä johti yhden mallin hylkäämiseen. Loput 32 mallia jaettiin arvioinnin tuloksen perusteella kolmeen laatuluokkaan. Laadittaessa Suomen tai sen lähialueiden ilmastoennusteita mallitietojen käyttäjät voivat hyödyntää tätä luokitusta valitessaan itselleen malleista sopivan osajoukon. Suomen vuotuiset keskilämpötilan ja sademäärän muutosarviot kylläkin osoittautuivat lähes samoiksi riippumatta siitä, käytetäänkö ennustusta laadittaessa koko mallijoukkoa vai jätetäänkö alhaisimmat pisteet saaneet mallit pois. Auringon säteilyn ennusteet olivat hieman herkempiä.

Julkaisijayksikkö: Sään ja ilmastonmuutoksen vaikutustutkimus (SIV)

Luokitus (UDK)	Asiasanat
551.581.1, 551.58, 551.583,551.588.7, 519.673	CMIP6-mallit, SSP-skenaariot, mallien laadunarviointi, MCPI-indeksi, ERA-Interim-uusanalyysit, Pohjois-Eurooppa
ISSN ja avainnimeke: 0782-6079 Raportteja	ISBN 978-952-336-141-6
DOI https://doi.org/10.35614/isbn.9789523361416	Kieli Englanti (tiivistelmä myös suomeksi ja ruotsiksi) Sivumäärä 48

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Utgivare	**Meteorologiska institutet**	Publikationens serie och nummer
	(Erik Palméns plats 1)	Rapporter 2021:7
	PB 503, 00101 Helsingfors	Datum 16.9.2021

Författare	Projektnamn
Kimmo Ruosteenoja	HEATCLIM, CHAMPS, FINSCAPES, LEGITIMACY
ORCID iD 0000-0002-4370-0782	

Rubrik: Globala CMIP6-klimatmodellers lämplighet för att utarbeta klimatscenarier i norra Europa

Sammandrag

I denna rapport har vi utvärderat prestantionsförmågan av nästan 40 globala CMIP6-klimatmodeller (CMIP6 = Coupled Model Intercomparison Project, fas 6). Fokusen ligger på det nordeuropeiska området, men förmogan att simulera sydeuropeiskt eller det globala klimatet har också studerats.

Modellutvärderingen inleddes med en teknisk kontroll genom att söka efter de absolut lägsta och högsta värdena. Ifall dessa värden visade sig vara orealistiska, övergavs utdatafilen eller hela modellkörningen. I detta skede övergavs en model helt och enstaka datafiler från två övriga modeller. Vid utvärderingen av de återstående modellerna var det främsta verktyget MCPI-indexet (MCPI =Model Climate Performance Index). MCPI-indexet kombinerar RMS-fel, som beräknats för de olika klimatvariablerna, till ett index. MPCI-indexet utforskar samtidigt hur nära modellresultaten och det observerade klimatet är vid olika årstider och vid områdets olika gridpunkter. MCPI-indexet beräknades för perioden 1981—2010 genom att jämföra modelresultat med ERA-Interimanalyserna. Klimatvariabler som undersöktes i utvärderingen var de följande: lufttemperatur, nederbörd, lufttrycket vid havsnivån och den inkommande solstrålningen på jordytan.

Förutom MCPI studerades RMS-felen i regionala medeltal separat för varje klimatvariabel och modellernas förmoga att simulera förändringarna hittils i den globala medeltemperaturen. Syftet var också att utforska ifall de globala temperaturförändringarna för olika växthusgasscenarier är förenliga med varandra. Dagens lägsta och högsta temperaturer undersöktes endast kvalitativt, eftersom uppgifter om dessa kvantiteter saknades från flera modeller. Dessa variabler ingick därför inte i den kvantitativa utvärderingen.

Fyra av de 37 modeller som klarat den preliminära tekniska kontrollen ansågs att vara helt oanvändbara för scenarioberäkningar. I två modeller var responser till de olika växthusgasscenarierna motsägelsefulla, och i två övriga modeller saknades data för en av de fyra viktigaste klimatvariablerna. För att minska beroendet mellan modellerna togs det dessutom inte med mer än två versioner av varje enskild model. Detta ledde till att en model övergavs. De återstående 32 modellerna delades in i tre kvalitetsklasser beroende på resultaten av utvärderingen. Användare av modelldata kan utnyttja denna klassificering för att välja passande antal av modeller som kan användas för att skapa klimatscenarier för Finland eller närstående områden. Årliga medeltemperatur- och nederbördsprognoser för Finland visade visserligen sig att vara nästan identiska oavsett om de härstammar från hela mängden av modeller eller om modeller som fått de lägsta utvärderingarna ignorerats. Solstrålning på jordytan var något mer känslig.

Publikationsenhet: Forskning av väder och klimatförändringens effekter

Klassificering (UDK)	Nyckelord
551.581.1, 551.58, 551.583,551.588.7, 519.673	CMIP6-modeller, SSP-scenarier, modellutvärdering, MCPI-indexet, ERA-Interim-analyserna, norra Europa
ISSN ja och serietitel: 0782-6079 Rapporter	ISBN 978-952-336-141-6
DOI https://doi.org/10.35614/isbn.9789523361416	Språk Engelska (sammandrag även på finska och svenska) Sidantal 48

Contents

1 Introduction

In projecting future climatic changes, numerical models constitute the primary tool (*IPCC*, 2013). Recently, a new generation of global climate models (GCMs), Phase 6 of the Coupled Model Intercomparison Project (CMIP6), has become available. The CMIP6 modelling effort is outlined in *Eyring et al.* (2016). In most cases, a more reliable picture of future climate and, in particular, the uncertainty of the projection is obtained by utilizing a large ensemble of models rather than a single or a few GCMs (e.g., *McSweeney and Jones*, 2016; *Stolpe et al.*, 2021). Nevertheless, prior to the calculation of projections, it is of utmost importance to evaluate the performance of the models. In this report, we assess how well CMIP6 GCMs can simulate the recent past mean climate and long-term observational trends. In addition, we check whether the future responses to divergent greenhouse gas scenarios are mutually consistent. Many of the present assessment methods are identical to those applied to the previous model generation (CMIP5) by *Luomaranta et al.* (2014).

The objective of the analysis is not to find any unambiguous ranking for the GCMs or to seek for the absolutely best-performing models. Rather, we aim at identifying those GCMs that manifest a low or even inadequate performance compared to the other models. In particular, it is essential to find out those GCMs for which the output data contain such fatal deficiencies that prevent utilization of the GCM in building future scenarios. For the remaining models, scores ranging from one to three stars are given. Such a scoring can assist users of the model data to decide what models to include in the analysis and what ones to abandon.

In order for the evaluation to be commensurable for all the GCMs, the procedure should utilize such climate variables for which data are available from every GCM. Therefore, any model that does not provide data for all the four key climate variables, namely surface temperature, precipitation, air pressure and solar radiation, will be disregarded immediately.

In the time when this assessment was performed (early 2021), the CMIP6 data archive was not yet complete. Additional parallel runs for the existing GCMs and, potentially, data from wholly fresh GCMs may emerge in the future. Accordingly, some calculations presented in this survey may then have to be updated.

The first step of the present model evaluation procedure consists of finding trivial technical faults by seeking the absolutely largest and smallest values of a climate quantity in the data files (section 2). Next, past trends in modelled global-mean temperatures are compared with observational estimates, and the consistency of future responses to the various greenhouse gas scenarios is assessed (section 3). Thereafter, the modelled baseline period (1981–2010) climate is compared with observational analyses, considering both northern and southern Europe and the global domain (section 4). Daily minimum and maximum temperatures are missing from such a large share of GCMs that these variables cannot be used in quantitative model evaluation. Nonetheless, some qualitative discussion considering these variables is presented in section 5. The applicability of the individual GCMs, considering all the evaluation criteria simultaneously, is assessed in section 6. Final concluding remarks are presented in section 7. In addition, there are two appendices, the first one dealing with the simulations of the EC-Earth3 model in more detail and the second one examining an alternative GCM evaluation tool GCMeval.

2 Downloading and preprocessing the model output files

Between November of 2020 and January of 2021, output files of 38 CMIP6 GCMs were downloaded from the data archives of the Earth System Grid Federation and stored on the Puhti computer hosted by the Centre for Scientific Computing (CSC), Finland. However, as will be detailed in section 2.1, the output files of one GCM turned out to be so badly corrupted that they had to be abandoned offhandedly. At this stage, monthly averages of the following variables were downloaded: near-surface air temperature (tas), precipitation (pr), surface pressure reduced to the sea level (psl), surface solar radiation (rsds), near-surface relative humidity (hurs), and the daily maximum and minimum temperatures (tasmax and tasmin). Historical model runs cover the years 1850–2014. To create future climate scenarios, model runs corresponding to four greenhouse gas scenarios were retrieved: SSP1-2.6, SSP2-4.5, SSP3-7.0 and SSP5-8.5 (*O'Neill et al.*, 2016). The number at the end of the acronym represents the strength of the scenario; e.g., the SSP2-4.5 scenario corresponds to a radiative forcing of about 4.5 Wm^{-2} in 2100. The scenario runs range from 2015 to 2100, however, for two GCMs (CAMS-CSM1-0 and IITM-ESM) only until 2099.

Temperature and surface pressure data were available for all the GCMs (Table 1). Precipitation data were missing from the KIOST-ESM model and solar radiation from MCM-UA-1-0 and some parallel runs of GFDL-ESM4. Relative humidity and daily extreme temperatures, by contrast, were lacking from quite a number of models.

Most GCMs provide data for several parallel runs: e.g., for the MIROC6 model, the count of parallel runs varies from 3 to 10, depending on the SSP scenario (Table 1)[1]. In the parallel runs, the temporal evolution of greenhouse gas concentrations and other forcing factors is identical, but initial conditions diverge across the runs. In general, we retrieved the results of all the parallel runs available, up to 10 runs (Table 1). However, for three GCMs (AWI-CM-1-1-MR, EC-Earth3 and EC-Earth3-Veg), the files had been given awkwardly in one-year snippets; in order to keep the workload reasonable, for these models a smaller number of parallel runs was examined. Fortunately, some of the output files of the two versions of the EC-Earth3 model were obtained directly in one piece through Kalle Nordling, a research scientist at FMI.

To facilitate the interpretation of the content of the dataset, we next examine an example model output file:

`tas_Amon_CNRM-CM6-1_ssp245_r4i1p1f2_gr_201501-210012.nc`

The name of the file contains the following information:

tas: the name of the variable (=surface air temperature)

Amon: monthly mean data (mon) belonging to the atmosphere realm (A)

CNRM-CM6-1: acronym of the model

ssp245: forcing scenario (SSP2-4.5)

r4: the number of parallel run (=4)

[1]For some models, the number of parallel runs available varies to some extent from one climate variable to another, and then the numbers given in Table 1 represent the mean temperature (tas) data.

i1: initialization index (=1)

p1: physics index (=1)

f2: forcing index (=2)

gr: grid information

201501-210012: temporal range (from January of 2015 to December of 2100)

nc: file format (NetCDF)

Table 1. Availability of the various climate quantities and the number of parallel runs for the different GCMs. Explanations: $*$: data for the quantity are available; $-$: data are missing completely; $(-)$: data are missing from so many model runs that it could not be used; D: data had to be rejected owing to technical errors; O: no attempt was made to retrieve the data. The five columns on the right show the count of parallel runs analyzed for the historical period (hist) and the four SSP scenarios (numbers; e.g., "126" refers to the SSP1-2.6 scenario). "tasnx" stands for the minimum and maximum temperature of the day.

	Model	Country	tas	pr	psl	rsds	hurs	tasnx	hist	126	245	370	585
1	MIROC6	Japan	*	*	*	*	*	*	10	10	3	3	10
2	MIROC-ES2L	Japan	*	*	*	*	*	*	10	3	1	1	1
3	MRI-ESM2-0	Japan	*	*	*	*	*	*	5	1	1	5	1
4	KACE-1-0-G	South Korea	*	*	*	*	*	(-)	2	3	3	3	3
5	KIOST-ESM	South Korea	*	-	*	*	*	-	1	1	1	-	1
6	TaiESM1	China (Taipei)	*	*	*	*	-	-	1	-	-	1	1
7	BCC-CSM2-MR	China (Peoples)	*	*	*	*	(-)	*	3	1	1	1	1
8	CAMS-CSM1-0	China (Peoples)	*	*	*	*	-	-	2	2	2	2	2
9	CIESM	China (Peoples)	*	*	*	*	-	O	3	1	1	-	1
10	FGOALS-f3-L	China (Peoples)	*	*	*	*	*	-	3	1	1	1	1
11	FGOALS-g3	China (Peoples)	*	*	*	*	*	*	6	4	4	4	4
12	NESM3	China (Peoples)	*	*	*	*	-	*	5	2	2	-	2
13	IITM-ESM	India	*	*	*	*	*	(-)	1	1	1	-	1
14	INM-CM4-8	Russia	*	*	*	*	*	*	1	1	1	1	1
15	INM-CM5-0	Russia	*	*	*	*	*	*	10	1	1	5	1
16	NorESM2-LM	Norway	*	*	*	*	*	-	3	1	3	1	1
17	NorESM2-MM	Norway	*	*	*	*	*	-	3	1	2	1	1
18	HadGEM3-GC31-LL	Britain	*	*	*	*	*	*	4	1	1	-	4
19	UKESM1-0-LL	Britain	*	*	*	*	*	*	7	10	5	10	5
20	MPI-ESM1-2-HR	Germany	*	*	*	*	*	*	10	2	2	10	2
21	MPI-ESM1-2-LR	Germany	*	*	*	*	*	*	10	10	10	10	10
22	AWI-CM-1-1-MR	Germany	*	*	*	*	(-)	*	3	1	1	3	1
23	CNRM-CM6-1	France	*	*	*	*	*	*	10	6	6	6	6
24	CNRM-CM6-1-HR	France	*	*	*	*	*	O	1	1	1	1	1
25	CNRM-ESM2-1	France	*	*	*	*	*	*	9	5	8	5	5
26	IPSL-CM6A-LR	France	*	*	*	*	*	*	10	5	7	10	5
27	CMCC-CM2-SR5	Italy	*	*	*	*	*	D	1	1	1	1	1
28	EC-Earth3	European union	*	*	*	*	*	*	6	4	6	4	4
29	EC-Earth3-Veg	European union	*	*	*	*	*	*	4	3	4	2	2
30	CESM2	United States	*	*	*	*	*	(-)	10	5	4	7	5
31	CESM2-WACCM	United States	*	*	*	*	*	(-)	3	1	5	1	5
32	GFDL-ESM4	United States	*	*	*	*	*	*	3	1	3	1	1
33	GISS-E2-1-G	United States	*	*	*	*	*	*	10	1	10	10	1
34	MCM-UA-1-0	United States	*	*	*	-	*	-	1	1	1	1	1
35	CanESM5	Canada	*	*	*	*	*	*	10	10	10	10	10
36	ACCESS-CM2	Australia	*	*	*	*	*	*	3	3	3	3	3
37	ACCESS-ESM1-5	Australia	*	*	*	*	*	*	10	10	10	10	10

The indices describing the initialization method and the physics of the model have a value of one (i1, p1) in all the files downloaded. This indicates that default options have been used for the initialization method and model physical parameterizations. However, the index characterizing the treatment of the climate change forcing differs from one in a few output files, i.e., the option f2 or f3 have been used. The f2 or f3 files were only used when data with the default option f1 were not available. For example, in the runs of the British UKESM1-0-LL and HadGEM3-GC31-LL models, the forcing index is 2 or 3, respectively. According to *Sellar et al.* (2020), this is mainly related to a different treatment of atmospheric ozone in the simulations. The impression of the author is that this does not have a substantial impact on the interpretation of the results of the scenario runs.

For the majority of models, the grid information parameter is "gn", indicating that the data are given on the original computing grid of the model. Alternatives "gr" and "gr1" tell us that some grid transformation has been performed before publishing the data. Evidently, that parameter is of minor importance when using the model data to compose climate change projections.

For some models, the results of the runs had already been given in a single file, covering either the years 1850–2014 (history runs) or 2015–2099/2100 (scenario runs). When this was not the case, the output data given in shorter time intervals were merged to a single file by using the "copy" command of the Climate Data Operators (CDO) software (https://code.mpimet.mpg.de/projects/cdo).

2.1 Preliminary technical check

To find simple technical data faults, the absolute minimum and maximum values were sought from every model output file (`file.nc`) by using the following CDO software commands:

```
cdo timmin -fldmin file.nc
```

and

```
cdo timmax -fldmax file.nc
```

This check covered all the climate variables listed in Table 1. The pre-inspection showed that the output files of the three historical runs (r1, r2, and r3) of the FIO-ESM-2-0 model were completely identical. The pr files were analysed comprehensively, and for the tas and psl files, we found the lowest and highest values to be the same in all the parallel runs. In addition, "checksum"-type warnings were repeatedly received when downloading the output files of the scenario runs. Therefore, the FIO-ESM-2-0 model had to be rejected altogether and is therefore not included in Table 1.

In the output files of CMCC-CM2-SR5, the daily minimum and maximum temperatures proved to be exactly identical to the daily mean temperatures, so these two variables had to be discarded from this model.

In the historical r3 run of KACE-1-0-G, the highest monthly-mean temperature was 90.02°C, the maximum monthly precipitation more than 9000 mm, and the lowest monthly mean temperature was not in line with the other parallel runs either. Hence, this parallel run was rejected, but the other two parallel runs of the model were included in the subsequent more detailed analysis.

In the CIESM model, both in the historical and SSP runs, precipitation was extremely low throughout, about three orders of magnitude lower than in observations. Apparently, the unit

of the quantity is different from that stated in documentation ($\mathrm{kgm^{-2}s^{-1}}$). By multiplying the precipitation data by 1000, fairly reasonable distributions were indeed obtained (for example, see Fig. 10), but there is still no guarantee that such a transformation would make the data correct.

Furthermore, three models (HadGEM3-GC31-LL, UKESM1-0-LL and ACCESS-CM2) produce huge monthly precipitation totals at individual grid points, at least 9000 or even about 35000 mm/month. However, this is not necessarily an indication of technical errors in the data, but such giant rains may rather result from issues in the model physics.

Near-surface relative humidity is a diagnostic output product of the model run that is not fed back into the model algorithm (*Ruosteenoja et al.*, 2017). Consequently, even unrealistic relative humidities do not necessarily make the model results questionable in other respects. Thus, no model was rejected owing to the strange behaviour of relative humidity, and in this report this quantity will not be used to assess model performances. Moreover, six GCMs do not provide any information on relative humidity (Table 1), which would make it difficult to use this quantity in the evaluation.

After the preliminary assessment, 37 out of the 38 GCMs were included in the subsequent phases of evaluation. The GCM output files having passed the technical check were next re-gridded onto a 2.5 x 2.5 degree grid covering the entire globe (73 x 144 grid points) by using the "remapcon" procedure of the CDO. The algorithm uses a method called "the first order conservative remapping". *Jones* (1999) states that, when transforming the data from a dense grid to a sparse one[2], it does not matter much whether the 1st or 2nd degree conservative remapping method is used, because the grid transformation then essentially consists of calculating spatial averages. Nevertheless, the lower order method consumes less computation time. In many other applications, however, the second-order method would be far more accurate (*Jones*, 1999).

When the source grid is far denser than the target grid, such a grid transformation that preserves the regional averages has pronounced benefits compared to, for example, linear interpolation. Linear interpolation only utilises those data points of the source grid that are closest to the target point, and the other grid point values do not have any influence on the result of the interpolation. Distortions occur especially when those few grid points are not representative for the entire target grid box, for example, if they happen to be located at high elevation compared to the remaining grid-square area.

3 Simulation of global mean temperature trends

3.1 Past trends

Figure 1 shows changes in the global mean temperature from the late 19th century to the turn of the millennium as derived from the historical runs of the CMIP6 models. According to observations, global mean temperature is estimated to have risen by 0.61°C (with an uncertainty range of 0.55–0.67°C) between the periods 1850–1900 and 1986–2005 (*IPCC*, 2013). The midpoint times of these periods are the same as those of the two 30-year periods examined in Fig. 1. In two GCMs (TaiESM1 and NorESM2-LM), global climate has warmed far too little, by

[2]For the majority of CMIP6 GCMs, the grid length is far smaller than 2.5°.

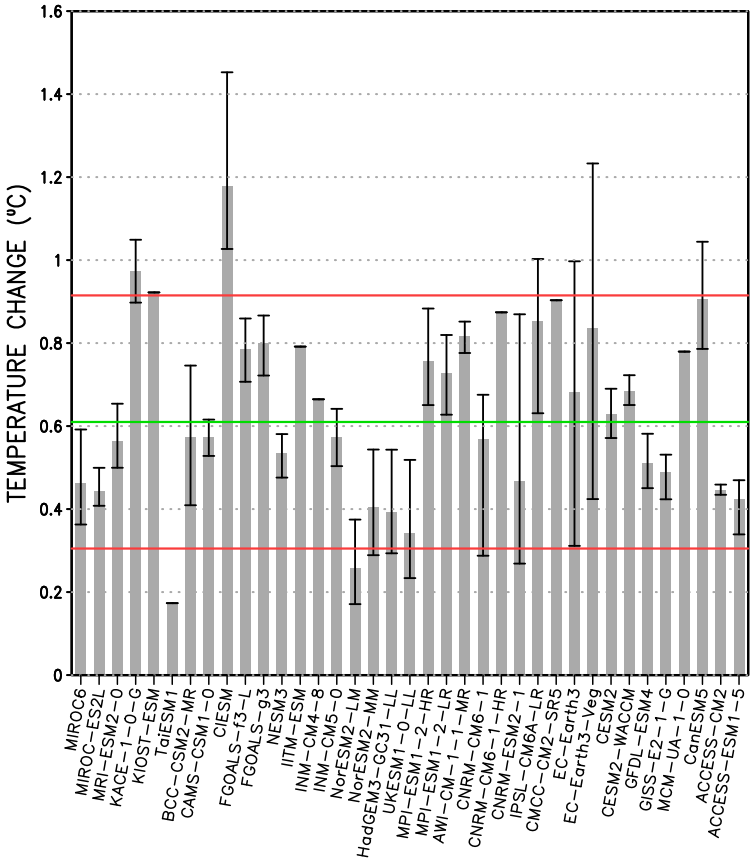


Figure 1. Changes in the global annual mean temperature from the period 1861–1890 to 1981–2010 as simulated by the various CMIP6 GCMs. Grey bars refer to the average of the parallel runs and black whiskers to the inter-realization range. The green horizontal line shows the corresponding observational change reported by *IPCC* (2013); for more details, see the text. Red lines represent that observational estimate multiplied by the factors of 0.5 and 1.5.

less than half of the observational estimate. Correspondingly, in three models (KACE-1-0-G, KIOST-ESM and, in particular, CIESM), the temperature has increased excessively, by more than one and a half times as much as what has been observed.

Furthermore, Fig. 1 shows that the simulated temperature increase varies quite considerably between the different parallel runs of the same model; typically, the differences are of the order of 0.2°C. This shows that natural variations even affect changes in the global mean temperatures quite substantially. Thus, it is not reasonable to require that the result of a single model run, or even the averages of several parallel runs, necessarily coincide very closely with the observed change. Accordingly, the discrepancy between the observed and model-produced warming makes doubtful only those five GCMs mentioned above.

Moreover, the magnitude of the differences between parallel runs varies greatly from one model to another. The differences are particularly large in both versions of the EC-Earth3 model; we shall re-examine this subject in Appendix 1 of this report. Since the number of parallel runs considered is fairly small, 1–10 per model, it is evident that stochastic factors affect the magnitude of the differences among the parallel runs substantially.

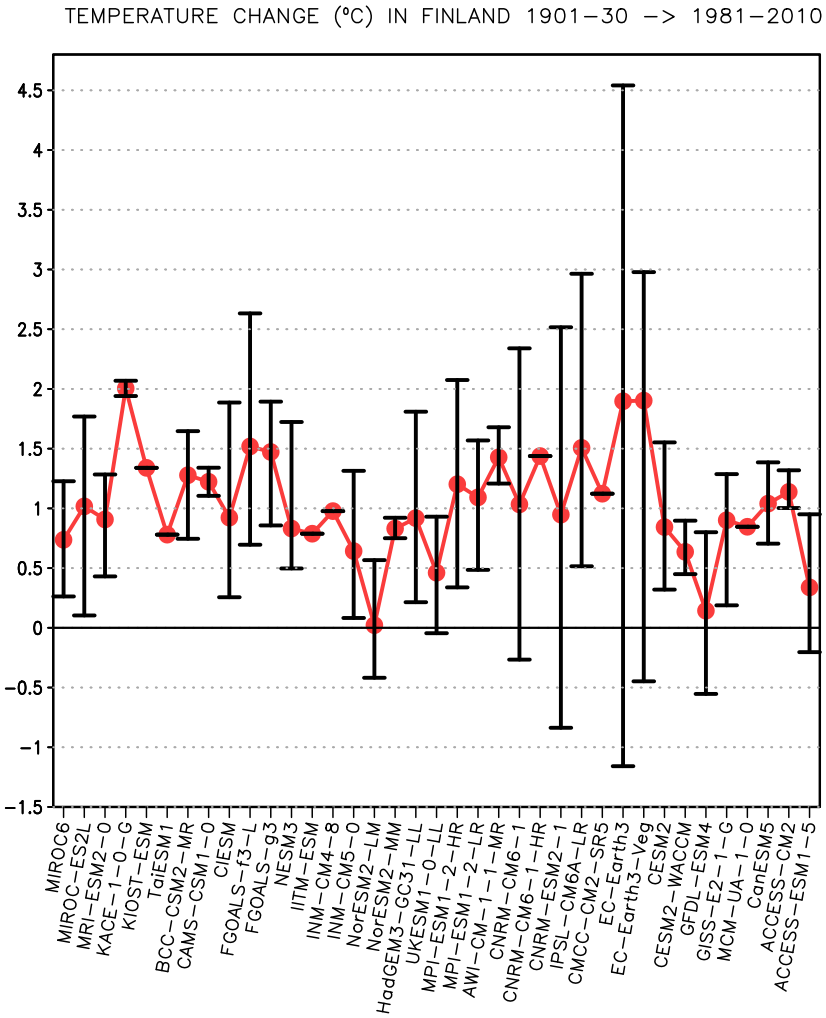


Figure 2. Changes in the annual mean temperature of Finland from the period 1901–1930 to 1981–2010 as simulated by the different models. The red line indicates the average of the parallel runs and vertical bars the inter-realization range.

In addition to changes in the global mean temperature, we studied changes in the average temperatures of Finland (Figure 2). It is evident that, for many models, the differences between the parallel runs are now very large: generally in the order of 1–2 degrees, but for the EC-Earth3 model as large as almost 6°C. Eight of the 37 models have at least one parallel run in which temperatures are even lower around the year 2000 than in the early 20th century. Likewise, there is a subset of eight models with some of the parallel runs simulating more than two degrees of warming.

Figure 2 indicates that in a region with such a small size as Finland, the contribution of natural variability in the simulated temperatures is very strong, even partially hiding the actual climate change signal. Therefore, we do not employ the compatibility of the modelled and observed Finnish temperature trends in dividing the models into goats and sheeps. In this respect, the procedure for evaluating CMIP5 models (*Luomaranta et al.*, 2014) diverged from the present one; in that work, one of the criteria of the performance was the concordance of modelled

Finnish temperature changes with observations.

3.2 Future changes — compatibility of different scenario runs

In order for climate change projections produced by a model to be usable, changes corresponding to the different SSP scenarios should relate reasonably to one another. Figure 3 shows, for each GCM and SSP scenario, the ratios of changes in the global mean temperature by the end of the century to the response to the SSP5-8.5 scenario. In general, these ratios are consistent among the models; for example, $\Delta T_{ssp245}/\Delta T_{ssp585} \approx 0.6$. Nevertheless, two notable exceptions from this rule draw attention. For the IITM-ESM model, the ratio $\Delta T_{ssp126}/\Delta T_{ssp585}$ is close to 0.8 rather than ∼0.4 as for the other GCMs. For CIESM, $\Delta T_{ssp245}/\Delta T_{ssp585}$ differs substantially from the general level.

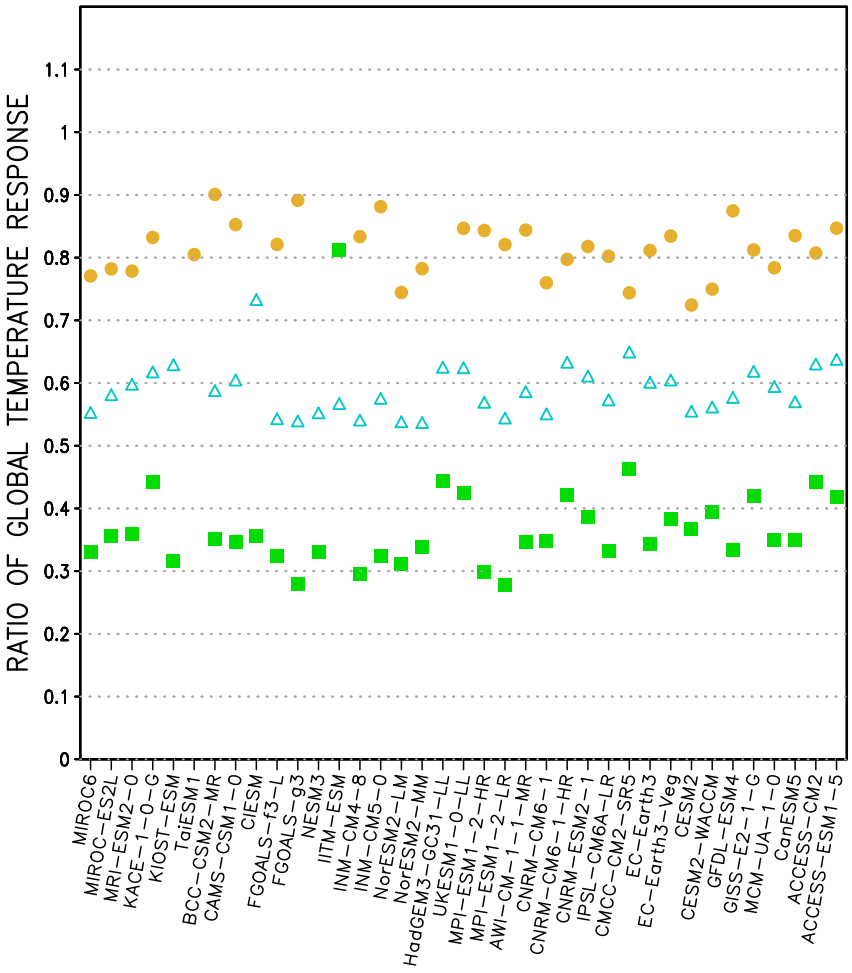


Figure 3. Changes in the global mean temperature (ΔT) from 1981–2010 to 2070–2099 in 37 CMIP6 GCMs expressed as the ratios of one SSP scenario to another. Green squares: $\mathbf{\Delta T_{ssp126}/\Delta T_{ssp585}}$; blue triangles: $\mathbf{\Delta T_{ssp245}/\Delta T_{ssp585}}$; yellow circles: $\mathbf{\Delta T_{ssp370}/\Delta T_{ssp585}}$. Here, $\mathbf{\Delta T}$ is calculated as an average of the parallel runs analyzed.

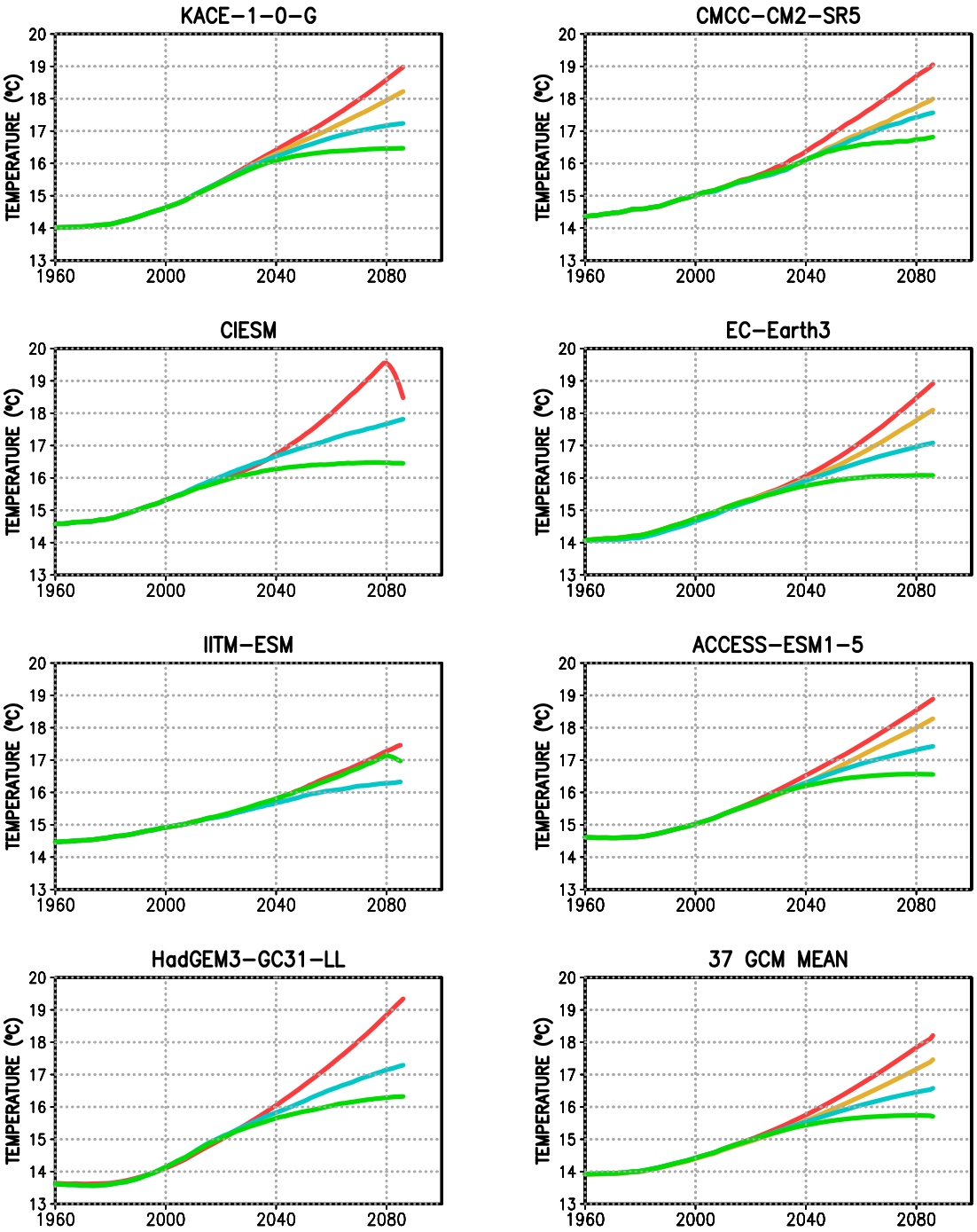


Figure 4. Time series of the global annual mean temperature for the period 1960–2086 simulated by seven CMIP6 models; 30-year running means: green - SSP1-2.6; blue - SSP2-4.5; amber - SSP3-7.0; red - SSP5-8.5. All time series are averages of parallel runs; the number of runs is shown in Table 1. No SSP3-7.0 simulations have been performed with CIESM, IITM-ESM and HadGEM3-GC31-LL. The lower right panel displays the corresponding time series of temperature as an average of the 37 GCMs.

Figure 4 shows the time series of global mean temperature from the 1960s to 2080s under different SSP scenarios as derived from the simulations of seven selected GCMs. For comparison, the corresponding multi-model means are given as well. For most models, changes corresponding to the different scenarios relate to one another in the same way as in the average of the results of the 37 GCMs. Nevertheless, in the SSP5-8.5 simulation of CIESM and the SSP1-2.6 run of IITM-ESM, the smoothed time series are peculiarly seen to manifest a steep decline after 2080. This behaviour is even more striking in the unsmoothed time series in which the drop takes place in the 2090s (not shown). In addition, in the IITM-ESM model the temperatures corresponding to SSP1-2.6 are strangely high. In fact, with the exception of the very end of the time series, they are close to the response to the SSP5-8.5 forcing scenario and much higher than those produced by the lower-emission SSP2-4.5 scenario.

It is therefore evident that these two models, CIESM and IITM-ESM, do not meet the condition of consistency among the responses to the different greenhouse gas scenarios. Other models do not appear to have any disorder in this regard.

4 Compatibility of the modelled and observed baseline-period climate

To evaluate the ability of the GCMs to simulate recent past climate, two main methods were applied. In both cases RMS differences between the modelled and observational monthly climatological means were calculated for the key climate variables. The basic metric utilized was the Model Climate Performance Index (MCPI) that combines RMS errors calculated for the different climate variables into one index (*Gleckler et al.*, 2008). In addition, we utilized a simplified performance index that examines the RMS errors separately for each variable (*Luomaranta et al.*, 2014).

Both indices were calculated over two regions, southern and northern Europe (Fig. 5). In addition, the MCPI index was determined for the entire globe. The simplified performance index ignores spatial variations within the domain and is therefore only suited for examining subcontinental or smaller scales. In comparing the model simulations and observational analyses, we intentionally examined larger areas than the territory of Finland alone, since in such a small area, natural variations have a strong impact on the 30-year climatological means (see Fig. 2 and related discussion).

4.1 Calculation of climatological means

For the calculation of the performance indices, long-term monthly averages for every climate quantity considered ($X =$ tas, pr, psl or rsds) are needed:

$$X_{t,ave} = \frac{1}{Y} \sum_{y=1}^{Y} X_{ty} \tag{1}$$

where X_{ty} stands for the mean value of a climate quantity for the month $t, t \in [1, 12]$ of the year y, and $Y = 30$ denotes the total number of the years. The time span used for the model validation consists of the years 1981–2010.

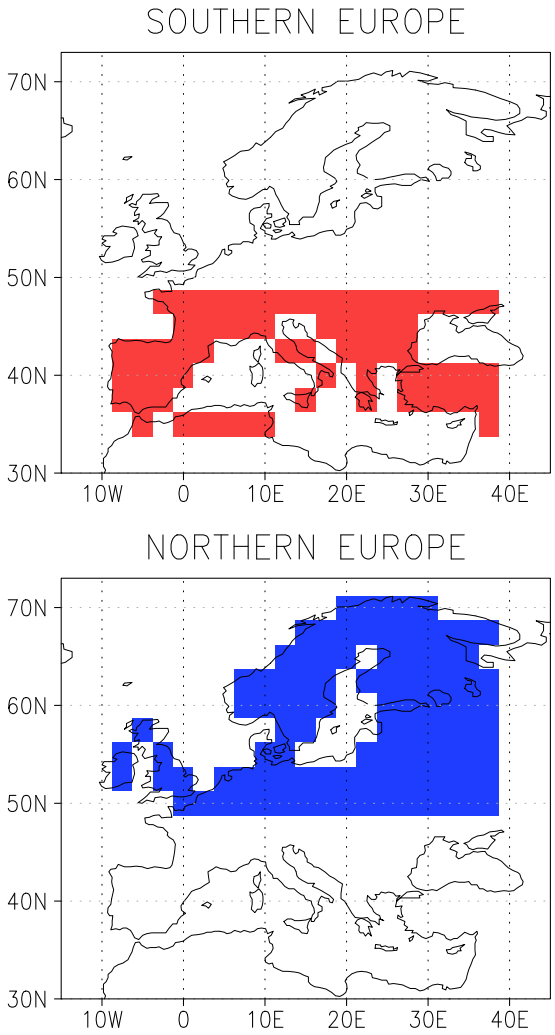


Figure 5. The southern and northern European domains applied in calculating the model performance indices, represented on a 2.5 × 2.5 degree grid.

Tridecadal climatological means were first calculated for all the individual GCM runs (Table 1), and these were used to calculate means over the parallel runs for every GCM. Observational climatological means used in the comparison were derived from the ERA-Interim re-analyses (*Dee et al.*, 2011). For tas and psl, we used the analysis fields. For pr and rsds, analysis fields were not available, and the ERA-Interim prediction fields were used instead.

4.2 The MCPI index

In calculating MCPI, the model performance is evaluated by studying RMS errors that consider deviations in both the seasonal cycle and the spatial distribution of the climate variable. *Gleckler et al.* (2008) stress that the comparison should not only use the annual means, but the seasonal cycle has definitely to be taken into account as well.

Firstly, one calculates the RMS difference between the modelled and observational distributions

of each climate variable included in the study by:

$$E_{\mu f} = \sqrt{\frac{1}{W} \sum_{\phi} \sum_{\lambda} \sum_{t} w_{\phi,\lambda,t} (F^{\mu f}_{\phi,\lambda,t} - R^{f}_{\phi,\lambda,t})^2} \tag{2}$$

where F denotes the modelled and R the observational (ERA-Interim) climatological monthly means, calculated by Eq. (1). W gives the sum of the weights $w_{\phi,\lambda,t}$.

Indices: ϕ stands for the latitude, λ for the longitude and t represents the calendar month. The index μ refers to the model or model run and f to the climate variable under consideration.

MCPI was calculated separately for three areas, namely southern and northern Europe (Fig. 5) and the global domain. For spatial weighting, Eq. (2) uses the relative surface areas of the grid squares, which are approximately proportional to $\cos \phi$.

As stated above, (2) was used to calculate RMS errors for each model run μ separately. From these RMS errors, the averages of the parallel runs were calculated to obtain an estimate of the RMS error for each model ($m = 1, ...M$).

Next, the median of E_{mf} was determined from the manifold of the RMS errors $E_{mf}, m = 1, ...M$; the median is denoted by $\overline{E}_f$. Note that, to avoid a single outlier model with a very large error to be dominant, the median rather than average of the RMS errors of the different GCMs is used.

The RMS errors of the individual climate variables f are expressed in different units and are thus not mutually comparable. To tackle this issue, normalized RMS errors were calculated by:

$$E'_{mf} = \frac{E_{mf} - \overline{E}_f}{\overline{E}_f} \tag{3}$$

E'_{mf} is a dimensionless number that can be expressed in decimals or percentages. For models outperforming the median model, $E'_{mf} < 0$, for models performing less successfully, $E'_{mf} > 0$. For example, if a certain model produces a RMS error that is 30 % smaller than the median among the models, $E'_{mf} = -0.3$. Accordingly, the relative RMS error provides an easily-understandable picture about how any individual model compares in performance to the remaining GCMs (e.g., *Gleckler et al.*, 2008, Fig. 3). Up to now, however, the climate variables have only been examined separately.

Finally, the Model Climate Performance Index (MCPI), which expresses the quality of simulations for multiple climate variables with a single numerical value, is calculated by:

$$MCPI_m = \frac{1}{N} \sum_{f=1}^{N} E'_{mf} \tag{4}$$

that is, an average of the relative RMS errors of the various climate variables is taken. Since the relative RMS errors are dimensionless numbers, (4) can be calculated even if the units of the quantities are different. In the equation, N denotes the total number of climate variables examined.

Gleckler et al. (2008) used a total of 26 climate variables to calculate the index, some of which describing conditions near the surface and others in the free atmosphere. However, in this study only four variables (tas, pr, psl, and rsds) have been utilized. No attempt was made to include

other variables (hurs, tasmin and tasmax) in the comparison, as data were missing from multiple GCMs (Table 1).

Conversely, for the four variables used, data were available from virtually the entire GCM ensemble, with two exceptions: pr was missing from the KIOST-ESM model and rsds from MCM-UA-1-0. Accordingly, for these two GCMs, MCPI could not be calculated. Nevertheless, as will be discussed in section 6, the absence of data for a key variable is regarded as a fatal deficiency for the model, and therefore the missing values of MCPI for these GCMs do not matter in practice; these two GCMs will be abandoned in any case.

To conclude, if a model or an individual model run has a positive value of MCPI, it is inferior to the "average" model according to this approach, while "good" models obtain negative MCPI values. Accordingly, just like in testing an infectious disease, it is a negative rather than a positive test result that is desired.

Figure 6 shows the values of MCPI for the various models, separately for northern and southern Europe and the global domain. For both half-continent sized areas, MCPI values of > 0.4 were regarded as suspicious. Looking at the average of the parallel runs, in the northern European subregion this threshold is exceeded by 5 GCMs (KACE-1-0-G, CIESM, FGOALS-f3-L, FGOALS-g3 and IITM-ESM) and in southern Europe by 3 GCMs (MIROC6, MIROC-ES2L and INM-CM4-8). Globally, the performance of the models appears to vary less widely, and a lower index value of 0.25 was selected for the alert limit. This threshold is exceeded by 3 GCMs (MIROC-ES2L, CIESM and IITM-ESM). It should be noted that, when examining the different areas, it is partly the same, partly different GCMs that fall in the high-risk group. MIROC-ES2L, CIESM and IITM-ESM have poor scores in two out of the three domains.

Finally, an unpleasant inference can be drawn from Fig. 6. If the regions from where the models originate (Table 1) are divided into two groups, one consisting of Asia, Australia and Russia and the other one of Europe and North America, one notes that in the former group the global MCPI is negative in only 3 out of the 16 GCMs (19%). In contrast, among the models coming from the latter group, 14 out of the 19 GCMs, or 74%, have an index with a minus sign. Of the five models that exceed the global suspicion threshold (MCPI>0.25), no one has been built in the latter group of continents[3]. In the opinion of the author, the success of the first-world GCMs is likely to be explained by the long traditions of academic education and development work of climate models and rather generous research resources. Another operant reason is the exploitation of less prosperous countries by attracting promising scholars and advanced scientists with financial resources and a favourable research environment. In this respect, Finland is not innocent either.

4.3 Simplified performance index

The compatibility of model results and observations can also be examined by using a simple performance index:

$$RMS_\mu = \sqrt{\frac{1}{12}\sum_{t=1}^{12}\left(\langle F^\mu(t)\rangle - \langle R(t)\rangle\right)^2} \tag{5}$$

[3]In fact, even in this group there was one rather badly-performing GCM, MCM-UA-1-0, but as stated above, no MCPI could be calculated for that model due to the missing rsds data.

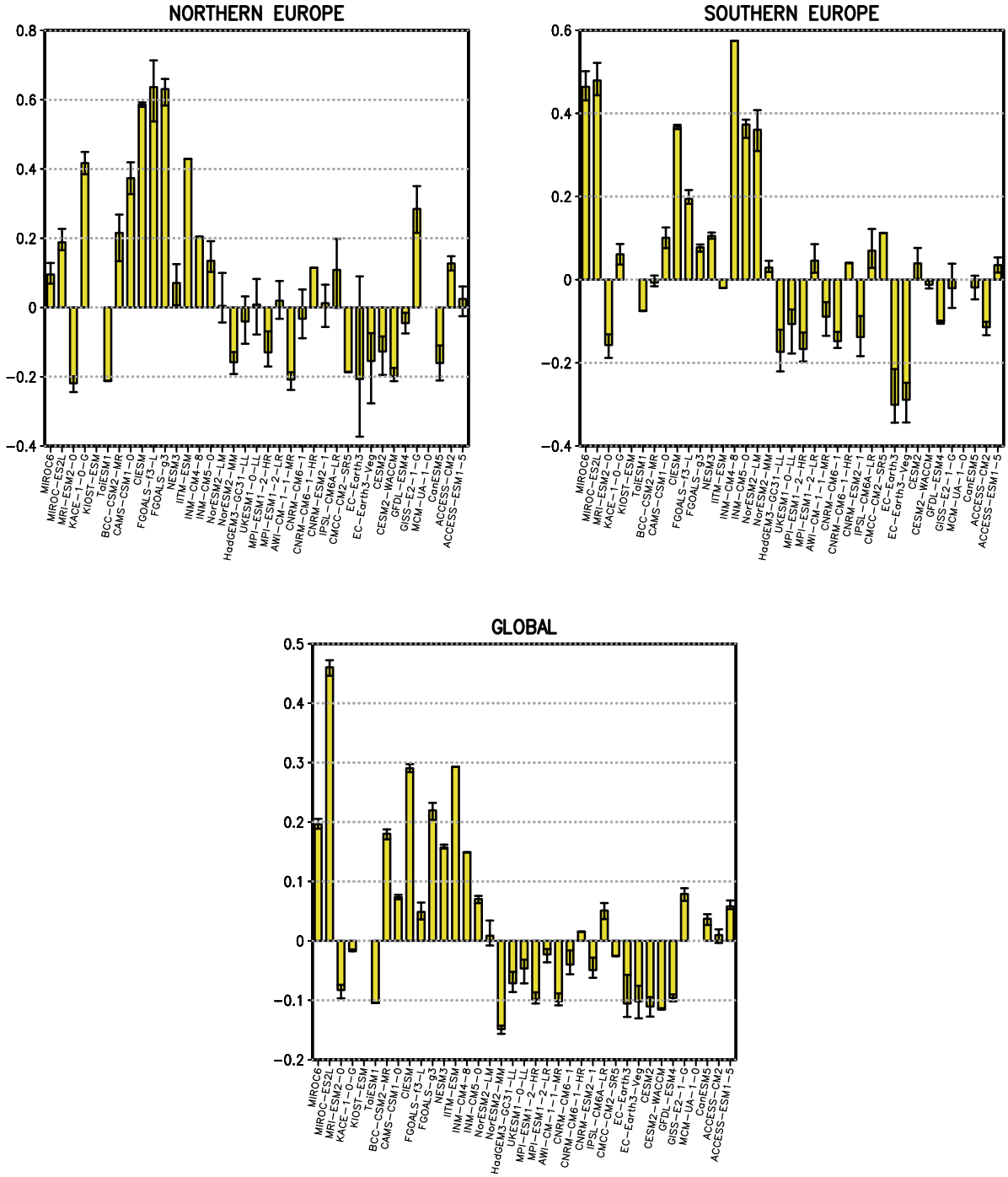


Figure 6. The values of MCPI derived from four climate variables for 35 GCMs (missing values for KIOST-ESM and MCM-UA-1-0), calculated for the regions of northern Europe (top left), southern Europe (top right) (the domains are defined in Fig. 5) and the entire world (bottom). The bars represent the averages of the parallel runs, whiskers the best- and worst-performing parallel runs of the GCM.

where $\langle\ \rangle$ represents the spatial average over the sub-region under consideration (southern or northern Europe, see Fig. 5). F is the model-derived monthly climatological mean of the period 1981–2010 and t the month. R represents the corresponding climatological mean calculated from the ERA-Interim analyses. The performance index was calculated separately for every model run μ, the number of parallel runs for all GCMs totalling 194 (Table 1). The simple index was calculated separately for all four climate variables: tas, pr, psl, and rsds. The index takes into account the seasonal cycle in the model results and observational analysis, but unlike MCPI, regional variations within the domain are disregarded. This entails both benefits and drawbacks. On the one hand, the index is somewhat less sophisticated than MCPI. On the other hand, in the simplified index small-scale differences between the modelled and observational fields do not have such a large importance as in MCPI. Large model versus re-analysis differences may occur, for example, in temperature fields in mountainous areas, since the limited resolution of the GCMs makes the representation of mountain ranges lower and more even than in reality (e.g. *Ruosteenoja et al.*, 2016b, Figure 1); the influence of such small-scale differences is partially cancelled out in the simple index. An noteworthy property of the simplified index is that the quality of the model simulation can be shown separately for all four climate variables.

A similar index was used in the evaluation of the CMIP5 models (*Luomaranta et al.*, 2014), but unlike in that work, now all the observational estimates used in the comparison have been derived from the ERA-Interim data.

In the northern European sub-region, the largest RMS differences from the re-analysis (Fig. 7) were found for the following GCMs: FGOALS-g3, KIOST-ESM and CAMS-CSM1-0 (temperature); CIESM, FGOALS-f3-L and KACE-1-0-G (precipitation); FGOALS-f3-L and IITM-ESM (surface pressure) and KIOST-ESM (solar radiation). In southern Europe, mean temperatures in MIROC6, precipitation in CIESM, surface pressures in both variants of the MIROC model and solar radiation simulated by both versions of INM proved to be of inferior quality (Fig. 8). However, the large precipitation bias in CIESM is likely to be explained (at least partially) by the fact that the unit reported for the quantity is incorrect — if the precipitation output produced by that model were multiplied by 1000, the error would be quite reasonable. In addition, in both versions of the EC-Earth3 model, the RMS errors of mean temperature produced by the different parallel runs diverge strongly, and the northern European temperature simulation in the lowest-performing run, r10, even falls near the lower end of the manifold of all the GCM runs. This topic is discussed in more detail in Appendix 1.

4.4 Qualitative comparison of model results with observations

Figure 9 shows the annual course of temperature, precipitation, surface pressure and solar radiation in some model runs that are relatively close to the re-analysis result, i.e., for which the MCPI index is definitely negative. All curves represent averages of the northern European region for the period 1981–2010. In these model runs, the deviations of the mean temperatures from the observational analysis are 0-3°C in winter and even smaller in summer. In surface pressure, the differences are likewise small, and with the exception of one GCM, this also holds for solar radiation; the AWI-CM-1-1-MR model tends to simulate too low radiation for spring. For precipitation, by contrast, errors of $\sim$20 % occur even in these well-performing model runs.

Correspondingly, Fig. 10 shows the seasonal course of the climatological averages of the four variables in model runs in which the northern European climate differs much from the reanaly-

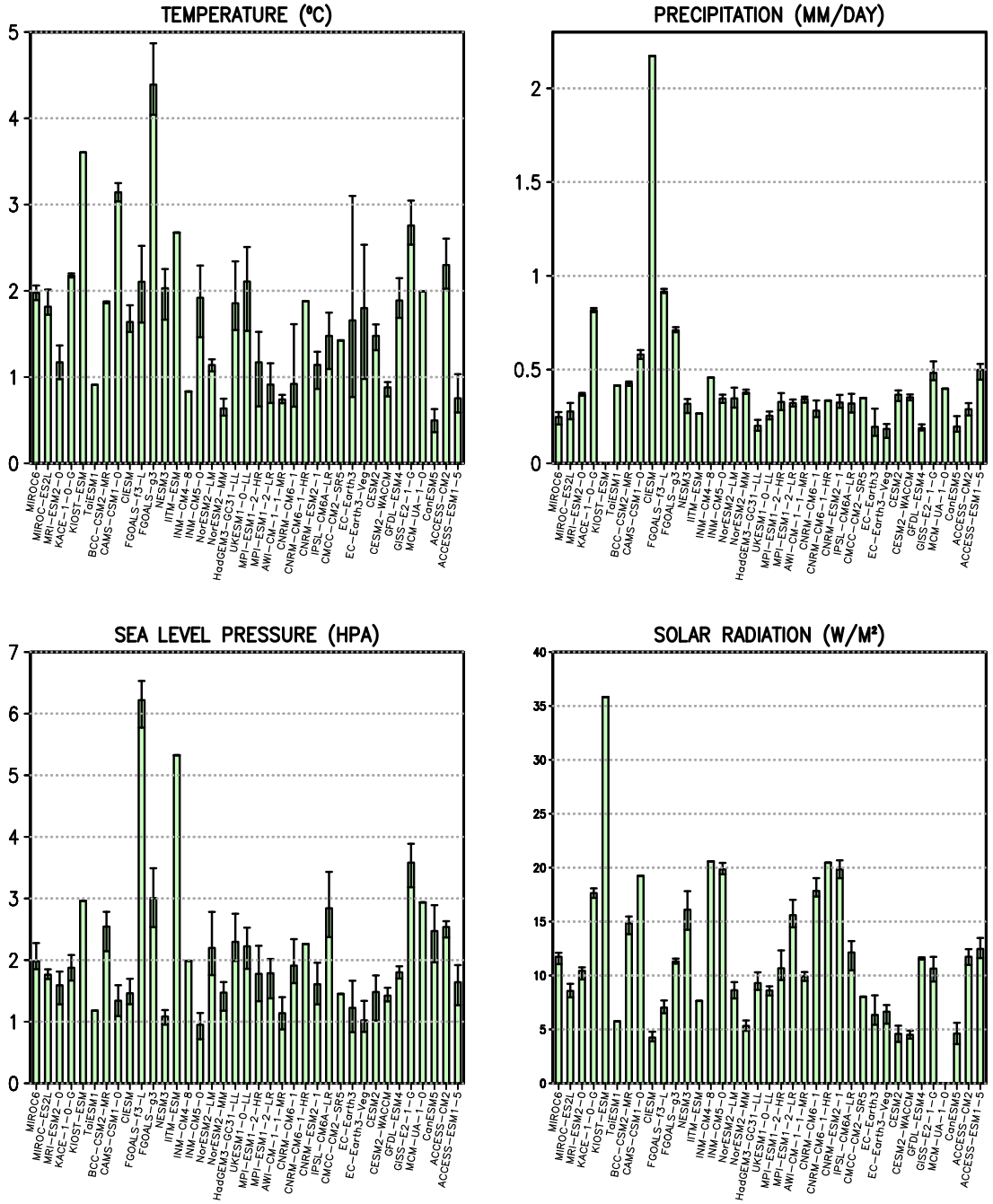


Figure 7. RMS differences between the simulated climatological means and those derived from the ERA-Interim analyses for the mean temperature, precipitation, surface pressure and solar radiation in northern Europe. Bars represent the average of the RMS errors in the different parallel runs of the GCM, whiskers RMS errors for the best- and worst-performing parallel runs. The KIOST-ESM model lacks the precipitation and MCM-UA-1-0 solar radiation data, i.e., for those cases RMS errors could not be calculated.

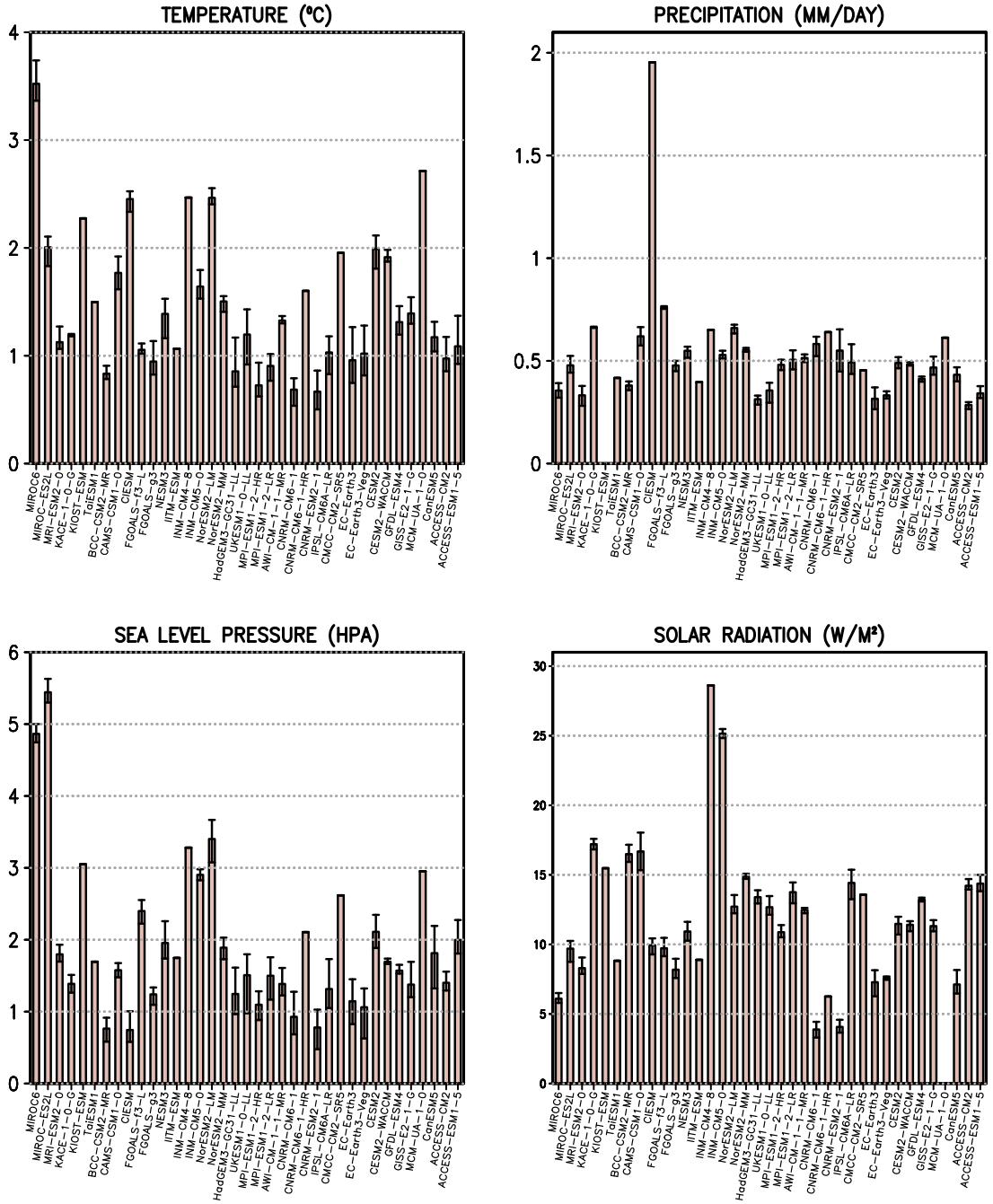


Figure 8. RMS differences between the modelled and ERA-Interim derived regional averages of mean temperature, precipitation, surface pressure and solar radiation in southern Europe. The notations are the same as in Fig. 7.

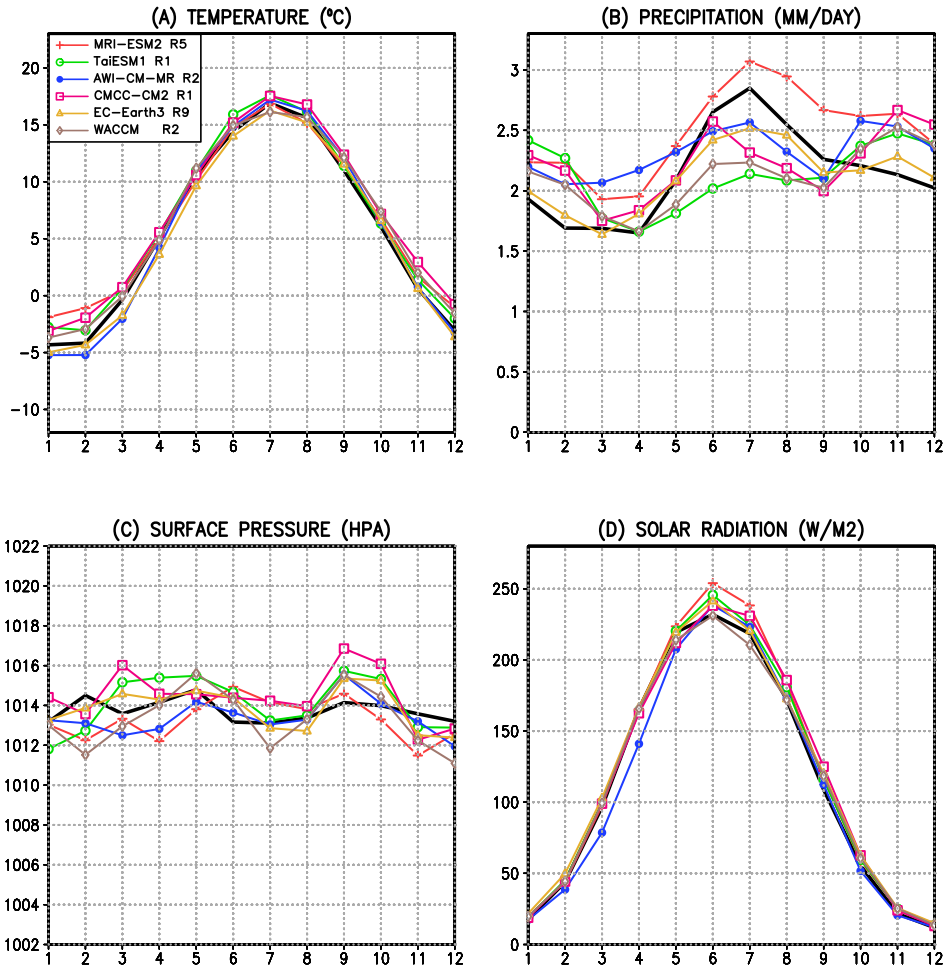


Figure 9. Annual course of (a) temperature, (b) precipitation, (c) surface pressure reduced to the sea level and (d) solar radiation in selected model runs classified as "good"; averages of the northern European sub-region for the period 1981–2010. The black curve without marks represents the climatological mean calculated from the ERA-Interim analysis, coloured curves the simulations of six GCM runs for which the MCPI index is clearly negative (the models are MRI-ESM2-0, TaiESM1, AWI-CM-1-1-MR, CMCC-CM2-SR5, EC-Earth3 and CESM2-WACCM; see the legend in the upper-left panel, where the indices of the parallel runs are also given).

ses. Now, monthly temperatures manifest negative biases up to 6°C and positive biases of $\sim$ 3°C. In some model runs, the phase of the seasonal distribution of precipitation is even opposite compared to the re-analysis. Also, air pressure tends to deviate from its observational counterpart much more than in the model runs shown in Fig. 9. Simulating the amount of solar radiation, by contrast, has been quite successful with the exception of two models (CAMS-CSM1-0 and KACE-1-0-G).

Figure 11 depicts the geographical distribution of the annual mean temperature bias for the same well-performing model runs that are considered in Fig. 9. In northern Europe, the deviations are generally smaller than two degrees, but far larger errors are seen in adjacent areas quite close to that region. The Barents Sea area in particular seems to be a difficult piece for some of these

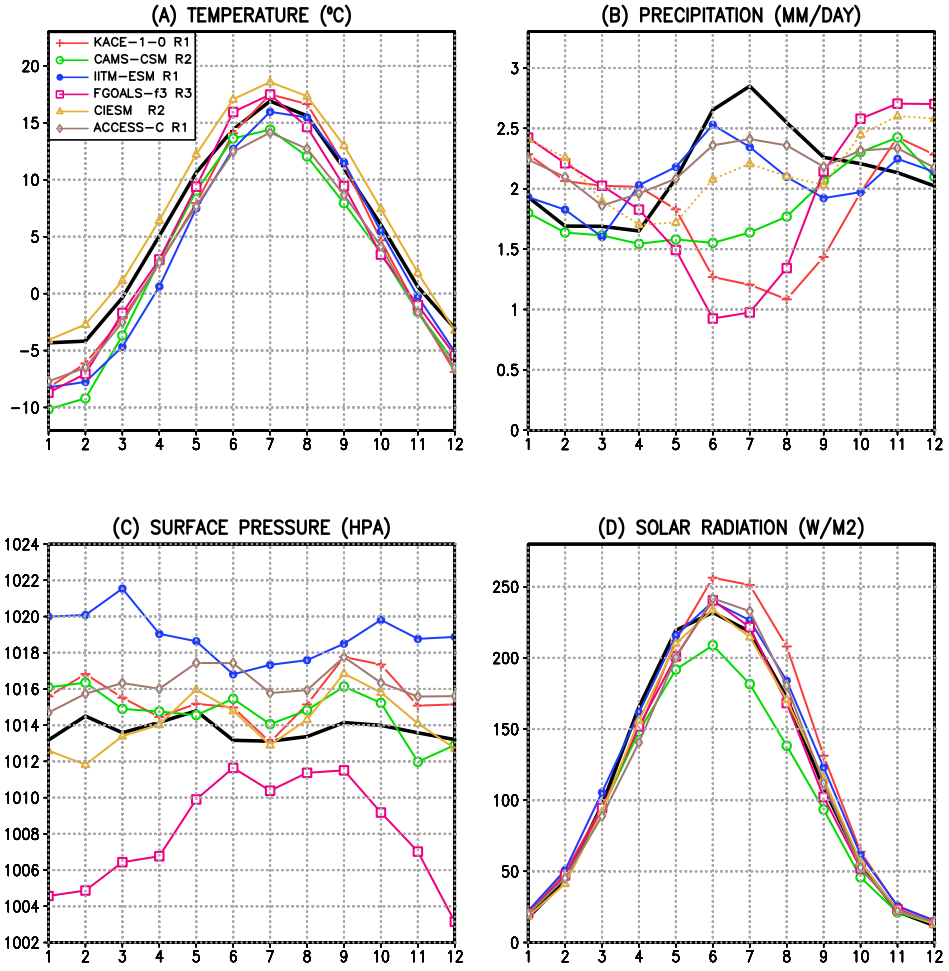


Figure 10. The annual course of (a) temperature, (b) precipitation, (c) surface pressure and (d) solar radiation in some model runs with a substantially positive MCPI index; see the caption of Fig. 9. The models are KACE-1-0-G, CAMS-CSM1-0, IITM-ESM, FGOALS-f3-L, CIESM and ACCESS-CM2; see the legend. According to the units reported in the model output files, the precipitation of the CIESM model would be negligibly small throughout the year. Nevertheless, the precipitation diagram also shows the precipitation produced by this model multiplied by 1000 (dark-yellow dashed line), although there is no compelling evidence that this is just the correct precipitation.

models.

Average annual temperature biases for selected model runs with a high MCPI index are shown in Fig. 12. Five of these models are far too cold, i.e., the average annual temperature in Finland, for instance, is about four degrees lower than in the ERA-Interim data. Conversely, in the CIESM model, the systematic error in temperature is fairly small. This shows that even if the simulated annual-mean temperature is relatively close to that observed, many other climate characteristics may be erroneous. The CIESM model suffers from deficiencies in the simulated precipitation, for instance.

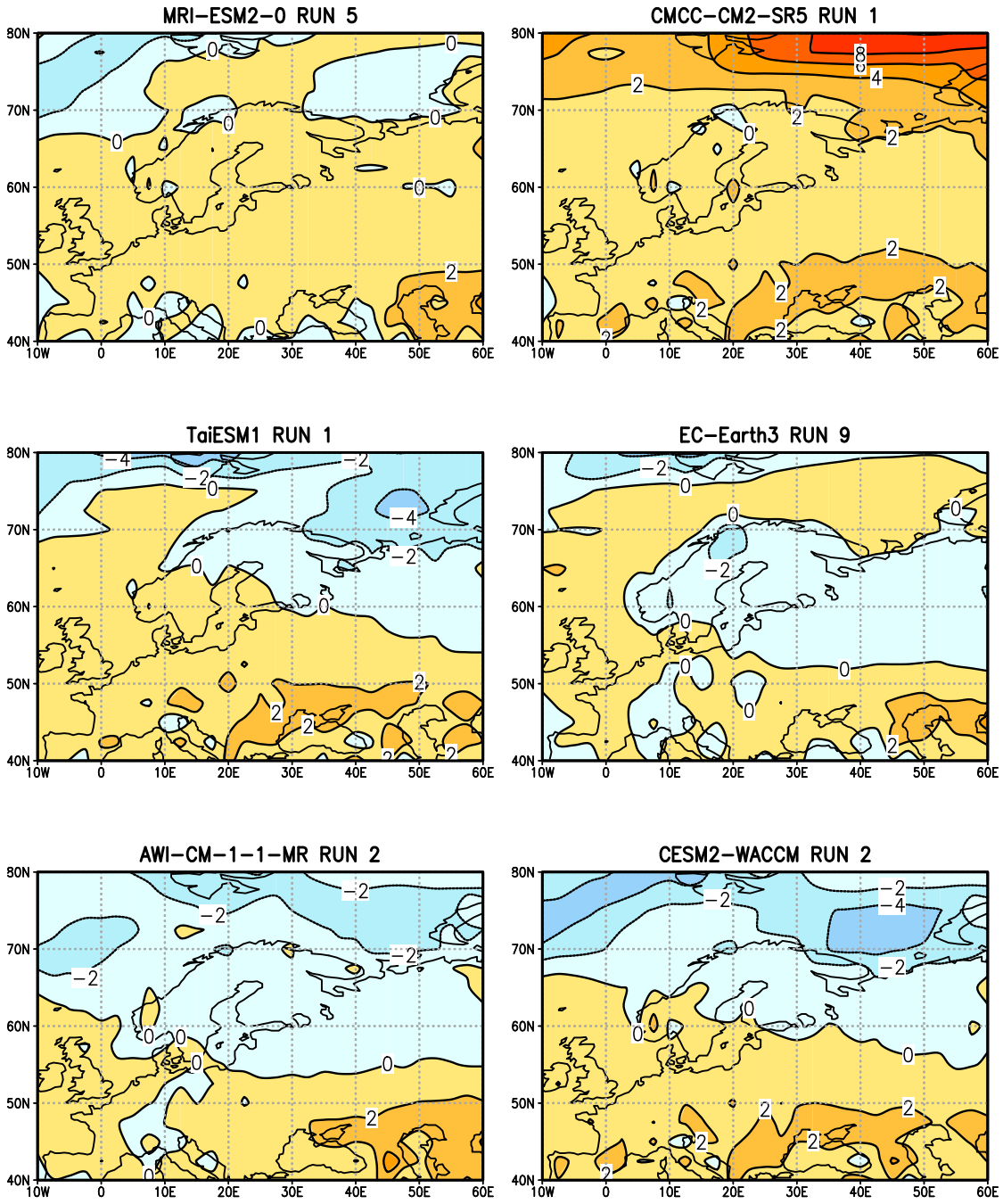


Figure 11. Deviations of the modelled annual mean temperature (in °C) from ERA-Interim in 1981–2010. The model runs are the same as in Fig. 9, i.e., the MCPI index is distinctly negative.

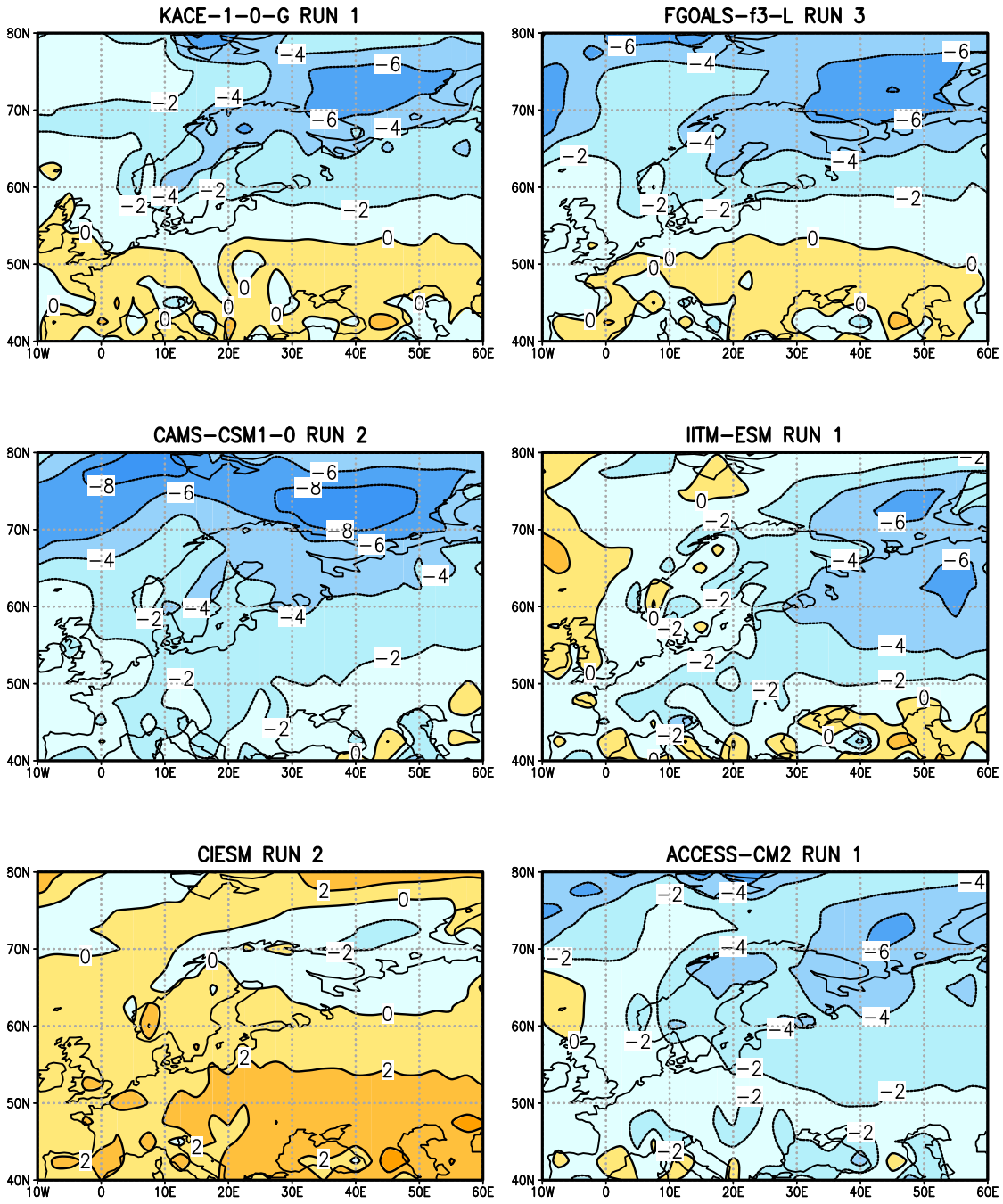


Figure 12. Deviations of the annual mean temperature (°C) from ERA-Interim in six model runs with a clearly positive MCPI index; the model runs the same as in Fig. 10.

5 Differences between the daily maximum and minimum temperatures

For the lowest and highest daily temperatures, data were only available from 24 of the 37 GCMs. Moreover, for one of them these data had to be ignored due to technical faults (section 2.1). It is difficult to compare these quantities with observations; for example, the daily minimum temperature in particular depends very much on local microclimate. In this report, we only compared the results of the different GCMs to one another to figure out whether some of the models produce quite unrealistic diurnal temperature amplitudes (Fig. 13).

It can be seen that in two models the daily temperature amplitude deviates much from the remaining GCMs: in NESM3 the difference is close to one degree throughout the year and in AWI-CM-1-1-LM it varies between 17–25°C depending on the season. Unrealistic diurnal temperature amplitudes occur in all months (Fig. 14) and over wide geographical areas (Fig. 15). In NESM3, the difference is close to one degree over continents and $< 0.5°$ over oceans. Conversely, in AWI-CM-1-1-MR the diurnal amplitudes are excessive. In February, for instance, even in ocean areas the amplitude is generally in the order of ten degrees or larger, and generally 20–30°C in the continental inland. For Finland in February, that model simulates average daily maxima that are near the freezing point while the minima are close to -26°C.

6 Scoring of the models

The deficiencies of the various GCMs, which were discussed in the previous sections, are summarized in Table 2. We first consider the minimum conditions that a GCM has to fulfil in order to be included in the scoring procedure:

- In order for the future projections calculated for the key climate variables to be consistent with each other, a model to be included in the scenario calculations must provide output data for the following four quantities: tas, pr, psl and rsds. The absence of precipitation and solar radiation data precludes the KIOST-ESM and MCM-UA-1-0 models, respectively (Table 1).

- If the simulated future global temperature change is not consistent across the different SSP scenarios, it is evident that the response to one or multiple SSP scenarios is erroneous. Then, it is impossible to use the model for projection calculations. This fatal error excludes the CIESM and IITM-ESM models (section 3.2).

- *Leduc et al.* (2016) have shown that different model versions developed by the same institution tend to produce more resembling results than GCMs in general. Consequently, to avoid giving excessive weight for any single modelling centre, it was decided here that no more than two versions of any GCM should be included in the calculations. For the CNRM model, three variants are provided (Table 1), from which CNRM-CM6-1-HR shows the weakest scores and is therefore excluded. This was done despite the fact that this model does not have any other minuses in Table 2. — It is somewhat surprising that the model variant having the highest resolution appeared to be the worst of the three versions. Perhaps the developers of the model have not been able to reconcile physical

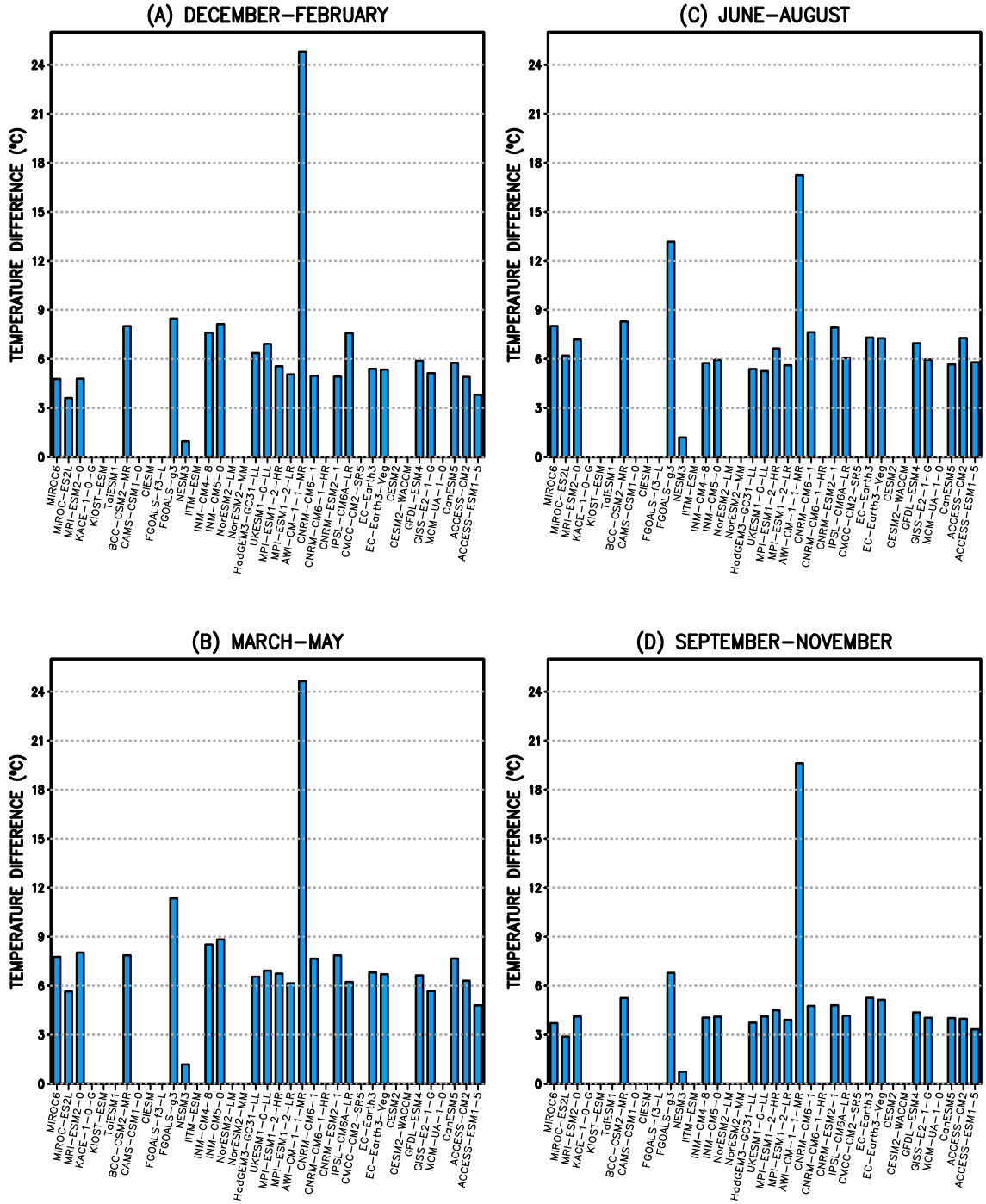


Figure 13. Average seasonal differences between the daily maximum and minimum temperature (°C) in 1981–2010 in different GCMs; spatial averages of Finland: (a) winter, (b) spring, (c) summer and (d) autumn.

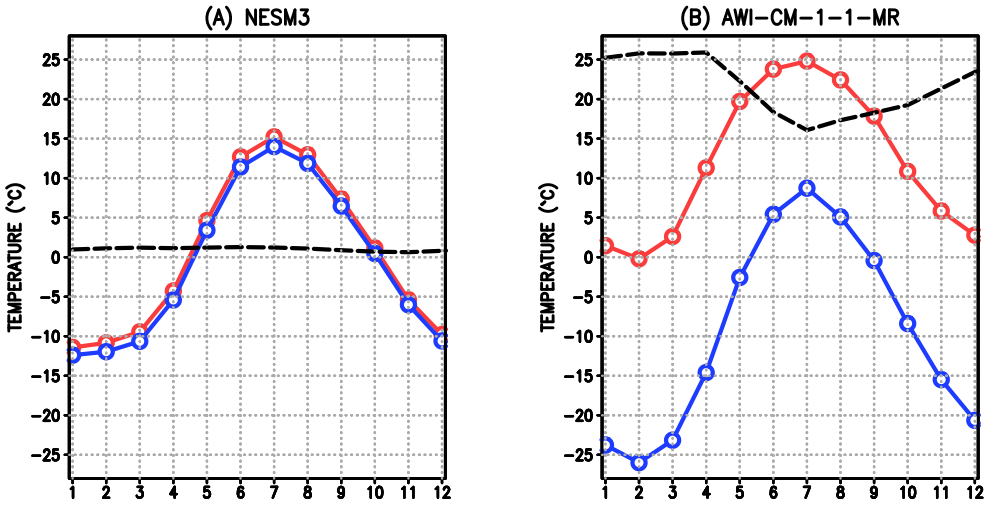


Figure 14. Annual course of the daily maximum (red) and minimum temperature (blue) and their difference (black dashed line) (°C) averaged over the period 1981–2010 in two badly-performing GCMs: (a) NESM3 and (b) AWI-CM-1-1-MR; spatial averages of Finland.

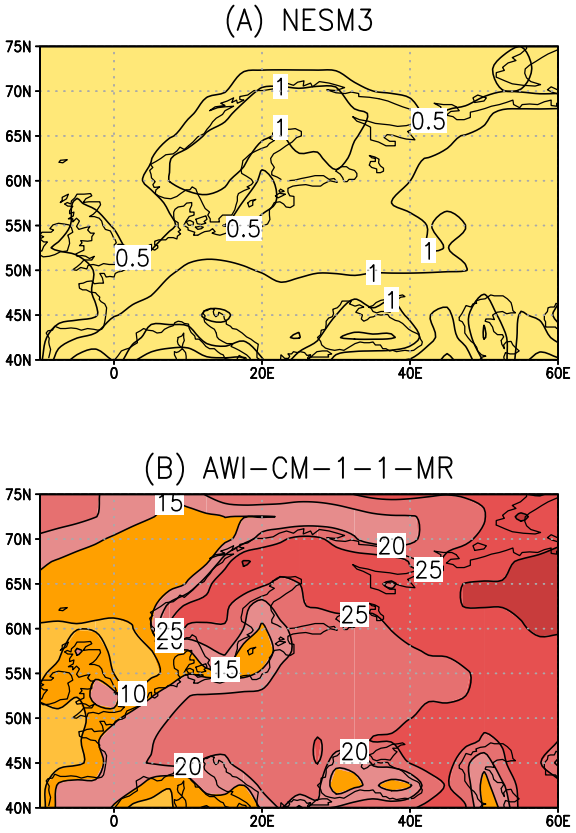


Figure 15. Temporal mean of the difference between the daily maximum and minimum temperatures for the period 1981–2010 in February simulated by (a) the NESM3 (substantial underestimation) and (b) AWI-CM-1-1-MR (large overestimation) model.

parameterizations with the high resolution; however, this topic is beyond the scope of the present report.

These five GCMs inevitably-excluded are marked by a capital "D" in Table 2. Otherwise, it is not possible to make any absolute decisions about the suitability of an individual model for climate scenario production, but performance assessment rather produces different shades of grey. The remaining GCMs were awarded scores using the following, admittedly fairly arbitrary formula:

$$\begin{aligned} P &= 25 - 100 \times (0.2 \times MCPI_{south} + 0.4 \times MCPI_{north} + 0.4 \times MCPI_{glob}) - 20 \times Trend \\ &\quad +5 \times Vers - 3 \times RMS - 3 \times Runs - 20 \times Precip \end{aligned} \tag{6}$$

Explanations and reasoning:

- MCPI indices (section 4.2) calculated for all three regions (southern and northern Europe and the global domain) are taken into account, but when simulating the Finnish or northern European climate, southern Europe is given a smaller weight.
- If the modelled past trend of the global mean temperature differs considerably from observations (section 3.1), the variable $Trend$ is given a value of 1; for CIESM the disagreement is so severe that $Trend = 2$. Otherwise no penalty is imposed, i.e., $Trend = 0$.
- If there are two versions of a model, the worse one is fined ($Vers = -1$) and the better of them is given a small credit ($Vers = +0.2$). The two British GCMs, HadGEM3-GC31-LL and UKESM1-0-LL, likewise bear great similarities (***Sellar et al.***, 2020) but have not been regarded as parallel model versions here. In any case, both models have received scores that are sufficient for three stars, regardless of whether this penalty/credit is given.
- If the RMS error calculated from the regional average (for at least one of the four variables; see section 4.3) is large for southern Europe, $RMS = 1$; if for northern Europe, then $RMS = 2$. If the error is large in both sub-regions, $RMS = 3$. However, this variable is given a fairly small weighting in the formula, since RMS errors derived from the regional averages to some extent provide the same information as the MCPI index.
- If the total number of available scenario runs is smaller than four, $Runs = 1$. In practice, this indicates that model runs are not available for all the SSP scenarios and even for the existing scenarios, there is a single run only (Table 1). The weight of this penalty factor is likewise small, for example, because the amount of model runs may increase in the future.
- If a model produces unrealistic precipitation totals (section 2), the variable $Precip$ is set to one, otherwise zero. Very heavy or small precipitation may be an indication of deficiencies in the model physics, and therefore quite a large emphasis is given to this variable.

In formula (6), the signs of the variables have been selected so that well-performing models obtain large positive scores. A negative score, on the other hand, is an indication of deficiencies. The constant of 25 at the outset of the formula helps to ensure that the signs of the scores are distributed conveniently among the models.

Table 2. Summary of the model performance assessment. Columns 1–2 show the model serial number and acronym. In columns 3–10, lower case letters indicate significant and upper case letters very serious problems with the model. "Variab": information on some key variable (tas, pr, psl or rsds) is missing (D); "Scen": simulated future global mean temperature changes under different SSPs are contradictory (D); "Vers": this is the worse of the two (x) or the worst of the three (D) versions of the model; "MCPI": MCPI has large positive values (s = southern Europe, n = northern Europe and g = global); "RMS": RMS errors of the regionally averaged climate variable(s) reveal significant differences from observations; "Trend": the simulated past global-mean temperature trend deviates significantly from observations (t or T); "Runs": less than four scenario runs are available (x; see Table 1); "Precip": the GCM output files contain either extremely small or large monthly precipitation totals somewhere in the world (x). In the last column, "D" indicates those models that have been considered necessary to be excluded from the scenario calculations. Other models have been given different performance ratings: "$***$" = no particularly serious deficiencies, fit to be used in scenario studies; "$**$" = significant shortcomings, inclusion should be contemplated; "$*$" = serious deficiencies, very strong consideration is necessary.

	Model	Variab	Scen	Vers	MCPI	RMS	Trend	Runs	Precip	Rating
1	MIROC6				s	s				***
2	MIROC-ES2L			x	sg	s				*
3	MRI-ESM2-0									***
4	KACE-1-0-G				n	n	t			*
5	KIOST-ESM	D			N.A.	n	t	x		D
6	TaiESM1						t	x		***
7	BCC-CSM2-MR									***
8	CAMS-CSM1-0					n				**
9	CIESM		D		ng	ns	T	x	x?	D
10	FGOALS-f3-L			x	n	n				*
11	FGOALS-g3				n	n				*
12	NESM3									***
13	IITM-ESM		D		ng	n		x		D
14	INM-CM4-8			x	s	s				**
15	INM-CM5-0					s				***
16	NorESM2-LM			x			t			**
17	NorESM2-MM									***
18	HadGEM3-GC31-LL								x	***
19	UKESM1-0-LL								x	***
20	MPI-ESM1-2-HR									***
21	MPI-ESM1-2-LR			x						***
22	AWI-CM-1-1-MR									***
23	CNRM-CM6-1									***
24	CNRM-CM6-1-HR			D						D
25	CNRM-ESM2-1									***
26	IPSL-CM6A-LR									***
27	CMCC-CM2-SR5									***
28	EC-Earth3									***
29	EC-Earth3-Veg			x						***
30	CESM2			x						***
31	CESM2-WACCM									***
32	GFDL-ESM4									***
33	GISS-E2-1-G									***
34	MCM-UA-1-0	D			N.A.					D
35	CanESM5									***
36	ACCESS-CM2			x					x	**
37	ACCESS-ESM1-5									***

If the value of P in (6) is ≥ 0, the model is considered not to have any serious weaknesses, and then that model is granted three stars. Accordingly, the three-star models are not necessarily brilliantly good, but a reasonably good performance is adequate for this score. The occurrence of deficiencies drops the number of stars to two ($-10 \geq P < 0$) or one ($P < -10$), the scoring thus depending on the number and severity of problem points.

The ratings given to the individual GCMs are, of course, fairly subjective. However, the same problem concerns all the model performance assessment methods developed. This issue is discussed further in section 7.

Accordingly, 5 of the 37 GCMs have definitely to be excluded from the scenario calculations. In Table 2, there are 24 models that have deserved three stars, 4 two-star and 4 one-star models. Thus, future scenarios might be calculated by using, for example, 24, 28, or 32 GCMs. For comparison, in studying the ensemble of CMIP5 models, *Luomaranta et al.* (2014) included 28 of the 35 candidate GCMs in the scenario calculations.

Furthermore, if two versions of the same GCM are available, that with the lower performance score may be abandoned, particularly if the score is low. For the FGOALS models, both versions have received only one star, but for FGOALS-g3 the number of parallel runs is substantially larger compared to FGOALS-f3-L (Table 1).

Admittedly, even GCMs originating from different institutions may bear similarities. Just like teenage girls borrow clothing from their friends, some GCMs have adopted sections of code from other GCMs. Furthermore, even if the codes have been programmed independently, similar parameterization methods are frequently employed in multiple GCMs (*Pennell and Reichler*, 2011).

For the AWI-CM-1-1-MR and NESM3 models, the differences between the minimum and maximum temperatures of the day were quite unrealistic (section 5). Therefore, these models should not be used in composing projections for those quantities. In other respects, these models appear to pass the tests and receive the full three stars (Table 2). Nevertheless, such a strange behaviour in extreme temperatures does make these models somewhat suspicious even in general.

6.1 Sensitivity of future projections to the size of the GCM ensemble

The sensitivity of multi-model mean future climate projections for Finland to the number of GCMs is studied in Fig. 16. Four alternative GCM sub-ensembles are explored: the entire set of 37 GCMs, those 32 GCMs that have received at least one star (Table 2), the 28 2–3-star GCMs and the 24 GCMs with three stars. Note, however, that simulations for SSP1-2.6 and SSP2-4.5 are missing from one GCM (TaiESM1) and those for SSP3-7.0 from 5 GCMs; in addition, in the 37-GCM ensemble precipitation and solar radiation data are both lacking from one GCM (Table 1). Hence, in those cases the responses are derived from a smaller number of models than the nominal count, and the responses to the different SSP scenarios are thus not wholly comparable. This may explain the qualitatively different temporal evolution of the solar radiation response to the various scenarios, for instance (Fig. 16).

For temperature and precipitation, the projections corresponding to the different number of models are nearly indiscernible, the differences ranging from zero to $\sim$0.2°C for temperature and to $\sim$0.6 percentage points for precipitation. Incident solar radiation appears to be somewhat more sensitive to the size of the model ensemble, with the largest differences of $\sim$1 percentage

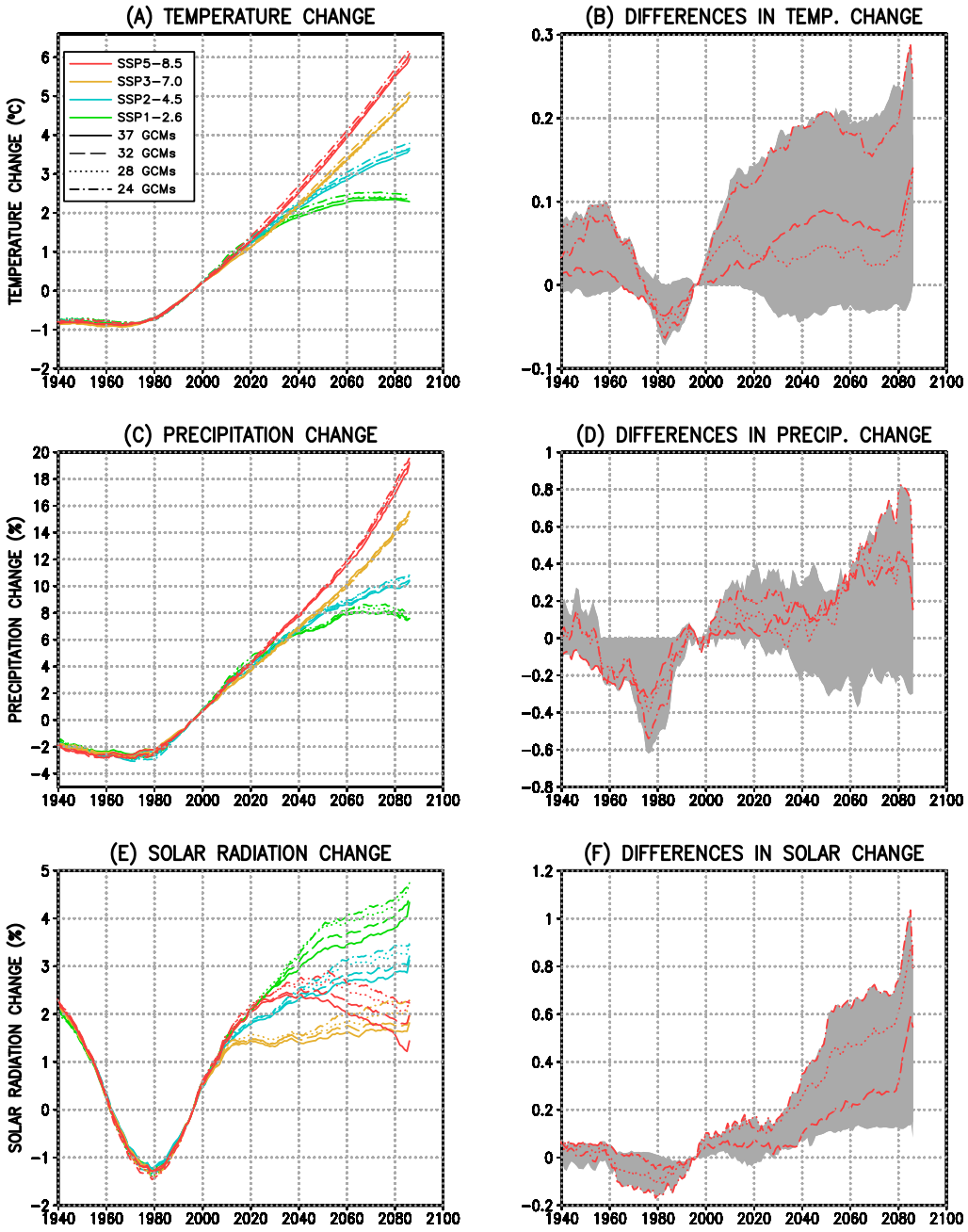


Figure 16. Left panels: Projected annual multi-model mean changes of (a) temperature (in °C), (c) precipitation (%) and (e) solar radiation (%) in 1940–2085; 30-year running-mean differences from the period 1981–2010, spatially averaged over Finland. Responses to SSP5-8.5, SSP3-7.0, SSP2-4.5 and SSP1-2.6 are denoted by red, amber, blue and green curves, respectively (see the legend). Correspondingly, the type of curve indicates the size of model ensemble: solid – all 37 GCMs; dashed – the 32 GCMs with at least one star in Table 2; dotted – the 28 GCMs with 2–3 stars; dash-dotted – the 24 GCMs having obtained three stars. Right panels: differences from the 37-GCM mean responses to SSP5-8.5 calculated by 32 (red dashed curves), 28 (dotted) or 24 (dash-dotted) GCMs for (b) temperature, (d) precipitation and (f) solar radiation. The full range of such differences considering all four SSP scenarios is shown by shading.

point. Moreover, the smaller the model ensemble, the more positive the projected solar radiation change. Even so, this may be a coincidence, since temperature and precipitation projections do not show such a monotonic dependence on the count of GCMs.

In Figs. 16b and f, the differences tend to increase sharply in the 2080s. This may be related to the strange behaviour of the CIESM and IITM-ESM models (Fig. 4). These GCMs are only included in the 37-GCM ensemble but not in the smaller sub-ensembles.

In any case, the main conclusion is that ignoring varying numbers of badly-behaving GCMs does not have any crucial impact on these simple projections. Even so, it is possible that inclusion of less qualified GCMs might more seriously distort projections for some other, more susceptible climate quantities.

Note that Fig. 16 is only intended to illuminate the sensitivity of the responses to the GCM ensemble size. No pattern scaling like that employed in *Ruosteenoja et al.* (2016a) has been utilized to create surrogate data for those SSP scenarios from which the actual model data is missing. Therefore, this exercise does not provide any "official" climate change scenarios for Finland.

7 Concluding remarks

In this survey, we originally inspected 38 GCMs, but one of them had to be abandoned owing to technical faults in the output data. In addition, four models proved to be inappropriate because of the lack of data for some key climate variable or discrepancies in future global warming between the different greenhouse gas scenarios. Moreover, to reduce inter-GCM dependencies, we did not include more than two model versions from any individual research centre. Because of this condition, one GCMs was left out.

The remaining 32 GCMs were divided into three categoria according to (i) their ability to simulate baseline period (1981–2010) climate in northern and southern Europe and globally; (ii) the consistency of past global-mean temperature trends with observations; (iii) the occurrence/absence of unrealistic precipitation totals and (iv) the count of available future scenario runs. Moreover, when two model version from the same research centre were available, the model version with a lower performance was penalized and the better one was granted a small credit.

It is quite evident that there does not exist any unambiguous method to evaluate the models. On the contrary, numerous alternative evaluation procedures have been developed (e.g., *Gleckler et al.*, 2008; *Flato et al.*, 2013; *Parding et al.*, 2020, and references in those papers), and they inevitably produce divergent ratings. Accordingly, the selection of evaluation criteria is always a more or less subjective matter. As an example, the GCMeval tool developed in Norway is discussed in Appendix 2 of this report.

For example, even the fairly widely-used MCPI index has been criticized by its authors in multiple ways (*Gleckler et al.*, 2008). Examples of issues:

- The MCPI index is regarded to be somewhat arbitrary. For example, there is no fundamental reason why all climate variables should be weighted equally when calculating the index (Eq. (4)).

- For each individual model, the relative RMS error may vary substantially across the climate quantities examined. Even if MCPI were fairly small, the error might be large for some individual quantity. In principle, this problem could be alleviated to some extent by assigning an enhanced weight to those climate variables that are deemed most important.
- Unfortunately, we do not yet know what are the particular features of observed climate that a model should be able to simulate well in order to be reliable in projecting future changes.
- MCPI only examines the concordance of long-term climatological means produced by the GCMs with observational data. However, in section 3.5 of *Gleckler et al.* (2008), the compatibility of temporal variability has also been discussed shortly. Even so, this option is not included in the index.

Accordingly, *Gleckler et al.* (2008) indeed do not praise MCPI as the final solution for ranking GCMs; it is not possible to infer with certainty how much the index actually tells about the reliability of models.

Besides MCPI, the RMS errors of seasonal variations in spatial means were utilized to assess the ability of the GCMs to simulate recent-past climate. If one studies very small domains, spatial variations in the modelled and observational fields ($F^{\mu f}$ and R^f) within the domain are minor. In that case, Eqs. (2) and (5) yield virtually the same RMS errors. As the dimensions of the domain are increased, spatial variations within the domain become increasingly important, and the outcomes of the two methods diverge. Accordingly, the simplified method is most appropriate for assessing model performance in sub-continental horizontal scale. Conversely, MCPI should be preferred in global scales. One useful property of the simple index is that the performance is assessed separately for the individual climate variables.

Model validation is also influenced by observational uncertainty. In particular, this concerns precipitation and solar radiation, for which the ERA-Interim data were derived from short-range forecasts of a weather model and not from the actual reanalyses, as was done for temperature and air pressure.

It is self-evident that we do not yet have any observational data from future climate that would allow direct verification of GCM projections. It is not clear how much the model performance in simulating recent-past climate and its trends tells about its ability to reliably simulate future changes.

Finally, it should be emphasized that the present GCM evaluation exercise is primarily valid for northern Europe. It is quite possible that a GCM having a low rating in Europe may be far better elsewhere in the world (or vice versa). Accordingly, the present findings should not be directly applied when assessing the suitability of the GCMs in other areas, particularly if the area considered is located far from northern Europe.

Acknowledgments

This work was supported financially by the Academy of Finland through the following projects: HEATCLIM (decisions number 329307), CHAMPS (329225), FINSCAPES (342561), LEGITIMACY (335562) and Atmosphere and Climate Competence Centre (337552). The CMIP6

GCM data were downloaded from the Earth System Grid Federation (ESGF) data archive (https://esgf-data.dkrz.de/search/cmip6-dkrz/). The modelling groups are acknowledged for making their model output available through ESGF. Computing resources were provided by the Centre for Scientific Computing (CSC), Finland. Kirsti Jylhä and Anna Luomaranta are thanked for critical comments and Antti Mäkelä and Lisa Haga for technical assistance.

References

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut and F. Vitart, 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137**, 553–597, doi:10.1002/qj.828.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer and K. E. Taylor, 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, **9**, 1937–1958, doi:10.5194/gmd-9-1937-2016.

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason and M. Rummukainen, 2013. Evaluation of climate models. In: T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 9, 741–866, Cambridge University Press, Cambridge, United Kingdom and New York, USA.

Gleckler, P. J., K. E. Taylor and C. Doutriaux, 2008. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, **113**, doi:10.1029/2007JD008972. D06104.

IPCC, 2013. Climate Change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, U.K., 1535 pp. [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)].

Jones, P. W., 1999. First- and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review*, **127**, 2204–2210, doi:10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2.

Leduc, M., R. Laprise, R. de Elía and L. Šeparović, 2016. Is institutional democracy a good proxy for model independence? *Journal of Climate*, **29**, 8301–8316, doi:10.1175/JCLI-D-15-0761.1.

Luomaranta, A., K. Ruosteenoja, K. Jylhä, H. Gregow, J. Haapala and A. Laaksonen, 2014. Multimodel estimates of the changes in the Baltic Sea ice cover during the present century. *Tellus A*, **66**, 22,617, doi:10.3402/tellusa.v66.22617.

McSweeney, C. F. and R. G. Jones, 2016. How representative is the spread of climate projections from the 5 CMIP5 GCMs used in ISI-MIP? *Climate Services*, **1**, 24–29, doi:10.1016/j.cliser.2016.02.001.

O'Neill, B. C., C. Tebaldi, D. P. van Vuuren, V. Eyring, P. Friedlingstein, G. Hurtt, R. Knutti, E. Kriegler, J.-F. Lamarque, J. Lowe, G. A. Meehl, R. Moss, K. Riahi and B. M. Sanderson, 2016. The Scenario Model Intercomparison Project (scenarioMIP) for CMIP6. *Geoscientific Model Development*, **9**, 3461–3482, doi:10.5194/gmd-9-3461-2016.

Parding, K. M., A. Dobler, C. F. McSweeney, O. A. Landgren, R. Benestad, H. B. Erlandsen, A. Mezghani, H. Gregow, O. Räty, E. Viktor, J. El Zohbi, O. B. Christensen and H. Loukos, 2020. GCMeval – An interactive tool for evaluation and selection of climate model ensembles. *Climate Services*, **18**, 100,167, doi:10.1016/j.cliser.2020.100167.

Pennell, C. and T. Reichler, 2011. On the effective number of climate models. *J. Climate*, **24**, 2358–2367, doi:10.1175/2010JCLI3814.1.

Ruosteenoja, K., K. Jylhä and M. Kämäräinen, 2016a. Climate projections for Finland under the RCP forcing scenarios. *Geophysica*, **51**, 17–50.

Ruosteenoja, K., J. Räisänen, A. Venäläinen and M. Kämäräinen, 2016b. Projections for the duration and degree days of the thermal growing season in Europe derived from CMIP5 model output. *International Journal of Climatology*, **36**, 3039–3055, doi:10.1002/joc.4535.

Ruosteenoja, K., K. Jylhä, J. Räisänen and A. Mäkelä, 2017. Surface air relative humidities spuriously exceeding 100% in CMIP5 model output and their impact on future projections. *Journal of Geophysical Research: Atmospheres*, **122**, 9557–9568, doi:10.1002/2017JD026909.

Sellar, A. A., J. Walton, C. G. Jones, R. Wood, N. L. Abraham, M. Andrejczuk, M. B. Andrews, T. Andrews, A. T. Archibald, L. de Mora, H. Dyson, M. Elkington, R. Ellis, P. Florek, P. Good, L. Gohar, S. Haddad, S. C. Hardiman, E. Hogan, A. Iwi, C. D. Jones, B. Johnson, D. I. Kelley, J. Kettleborough, J. R. Knight, M. O. Köhler, T. Kuhlbrodt, S. Liddicoat, I. Linova-Pavlova, M. S. Mizielinski, O. Morgenstern, J. Mulcahy, E. Neininger, F. M. O'Connor, R. Petrie, J. Ridley, J.-C. Rioual, M. Roberts, E. Robertson, S. Rumbold, J. Seddon, H. Shepherd, S. Shim, A. Stephens, J. C. Teixiera, Y. Tang, J. Williams, A. Wiltshire and P. T. Griffiths, 2020. Implementation of U.K. Earth System Models for CMIP6. *Journal of Advances in Modeling Earth Systems*, **12**, e2019MS001,946, doi:10.1029/2019MS001946.

Stolpe, M. B., K. Cowtan, I. Medhaug and R. Knutti, 2021. Pacific variability reconciles observed and modelled global mean temperature increase since 1950. *Climate Dynamics*, **56**, 613–634, doi:10.1007/s00382-020-05493-y.

Appendix 1: Special issues of the EC-Earth3 model

In the main text of the report, the special nature of EC-Earth3 compared to most other GCMs has already been noted:

- When looking at the averages of the parallel runs, EC-Earth3 performs quite nicely: past trends in the global mean temperature are reasonably consistent with observations (Fig. 1); future responses to the different SSP scenarios fit well together (Fig. 3); the value of the MCPI index is one of the best among the ensemble of models, and in southern Europe the two versions of the model even occupy both top positions (Fig. 6); especially the simulation of precipitation seems to succeed well (Fig. 7).
- On the other hand, differences between the various parallel runs, both in the past trends and in the quality of the baseline-period climate, are very large in both versions of the

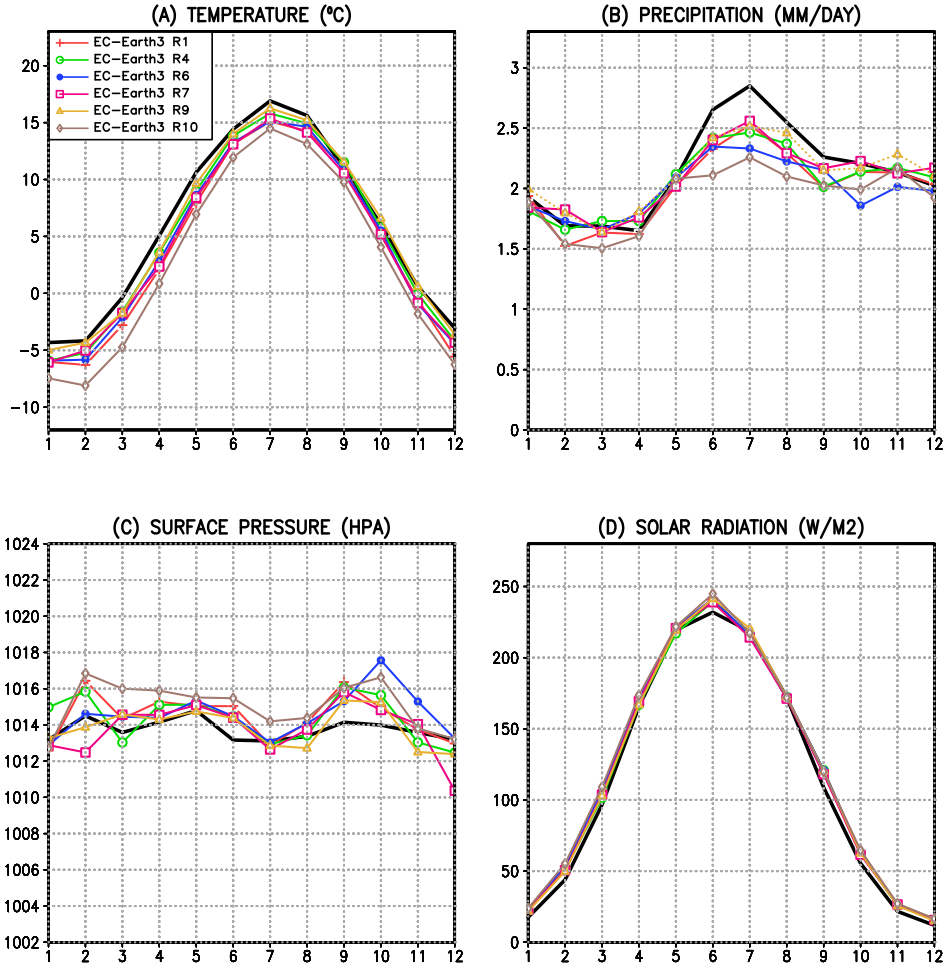


Figure 17. Annual course of the climatological (1981–2010) northern European regional mean (a) temperature, (b) precipitation, (c) surface pressure and (d) solar radiation in parallel runs 1, 4, 6, 7, 9 and 10 of the EC-EARTH3 model. For further information, see the caption of Fig. 9.

EC-Earth3 model, larger than in any other model. This indicates that even if the average of the parallel runs works well, some individual parallel runs would belong to the weakly-performing tail of the ensemble (Figs. 1, 6a and 7a).

The latter finding might also be interpreted as indicating that if, for some reason, only one parallel run were available from the EC-Earth3 model and this single run by chance were just the worst one, the model could even be threatened with a rejection when simulating the northern European climate. But fortunately for the model, six parallel runs have been analyzed, and thus the overall score is good.

Correspondingly, it is possible that some other GCM actually does have similar large differences between parallel runs, but the differences do not materialize because the number of parallel runs is too low, a single or a few ones only. Due to such a small number of parallel runs, the quality score of that model may therefore become distorted if those few runs by chance happen to fall very close to or very far from observed climate and its past trends.

The performance of the different parallel runs of EC-Earth3 has been studied in more detail in Figures 17–20. In examining seasonal variations in northern Europe (Fig. 17), the 10th parallel run differs sharply from the other ones: temperatures are generally lower than in the other runs, precipitation is likewise lower in most months and air pressure is higher. All of these factors act to divert this parallel run further away from observational climate. The other runs are also slightly too cool compared to the re-analysis and, in summer, too dry as well, but the differences both among these runs and compared to the re-analysis are smaller than those in the 10th run.

As regards the annual mean temperatures, the 10th parallel run differs from the other runs in Europe by being too cold (Figure 18). The difference is amplified north of the 60th latitude and especially in northern oceans; the annual means are up to 6–12°C too low. The contrast is particularly striking compared to the 9th run. In that run, biases in temperature are quite small throughout the area. Over ocean areas north of 70°N, temperature differences between these two parallel runs are widely in the order of ten degrees.

Figures 19 and 20 show the annual mean and summer-season differences in precipitation between the individual simulations of the EC-Earth3 model and the ERA-Interim analysis. The main features of the difference are the same in all parallel runs. In northern ocean areas and south-eastern Europe, the model produces too little precipitation on an annual basis, but in northern and western Europe the concordance is good. In summer, the most notable feature in the simulations is the aridity of south-eastern Europe. For example, around the Black Sea, in the model simulations precipitation is only about a third of what it should be. Compared to the other runs, the 10th parallel run is somewhat drier in northern Europe and its adjacent northern-ocean areas. However, given the large regional differences in the precipitation deviation, the dryness is far less striking than the coldness of that run (Figure 18).

Figure 21 shows how the annual mean temperature of Finland and the entire northern European area develops in different parallel runs of the EC-Earth3 model. During the 19th and 20th centuries, the differences among the parallel runs are large, at most 4–5 degrees in Finland, and somewhat smaller when looking at northern Europe as a whole. In addition, in all the parallel runs, temperatures fluctuate on a time scale of 40–60 years, in Finland by about three degrees. For example, in the 10th parallel run, the Finnish climate would have cooled from the 1920s to the 1980s by about 3°C, while in the 9th run in the same time, temperatures would have risen by almost 4°C. Temperatures tend to vary in phase in Finland and elsewhere in northern Europe;

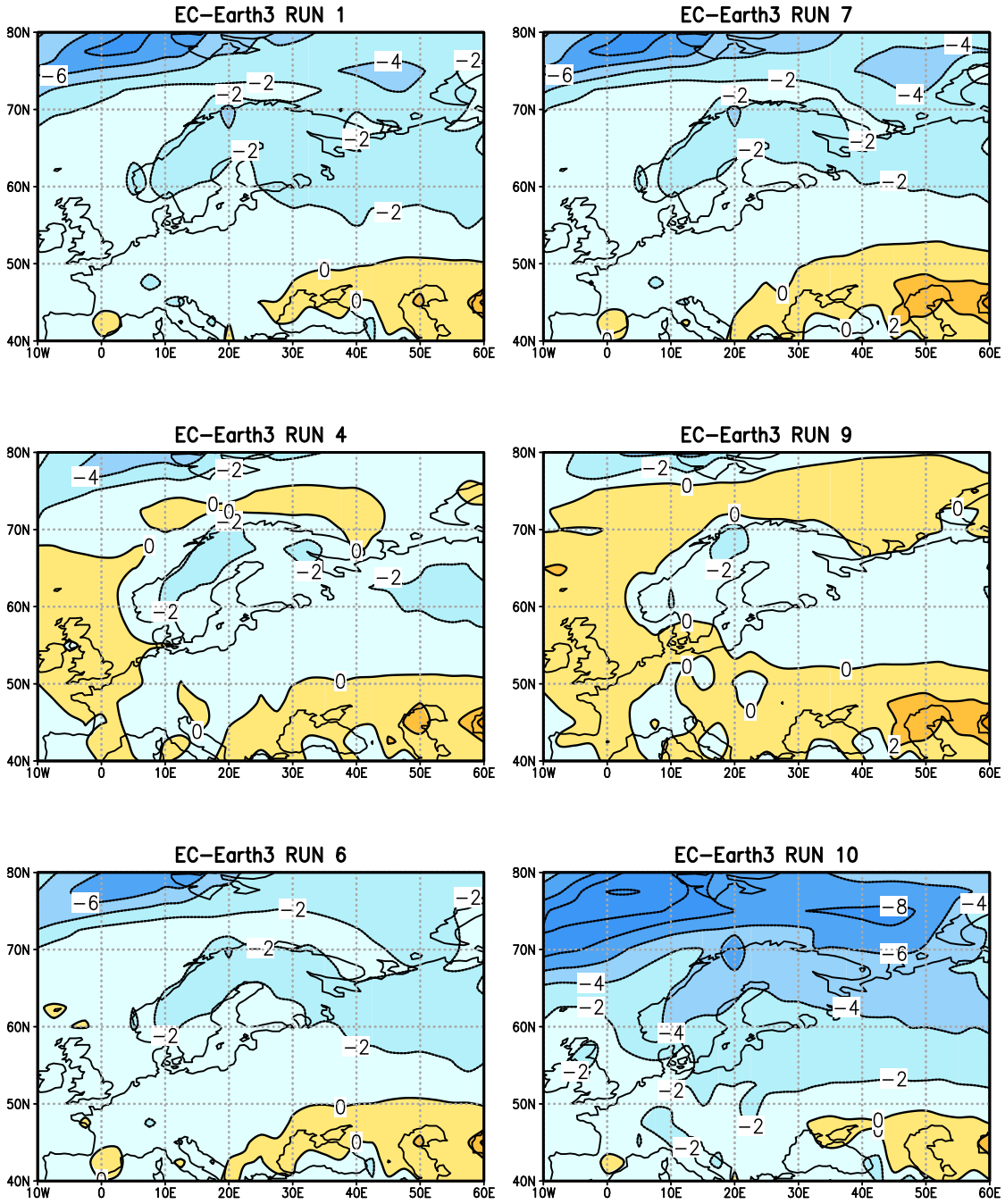


Figure 18. Deviation of the annual mean temperature (in °C) from the ERA-Interim re-analysis in six parallel runs of the EC-Earth3 model.

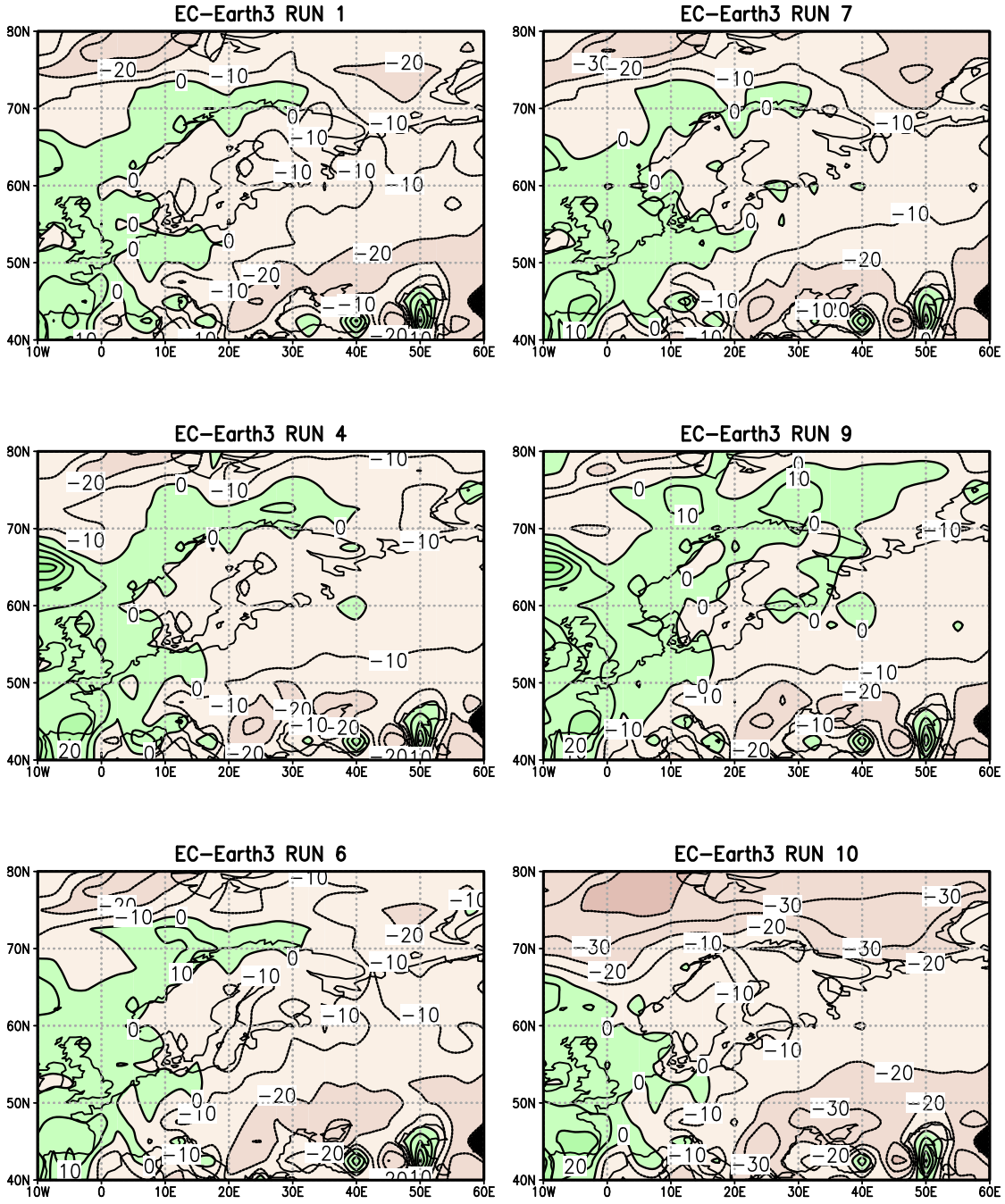


Figure 19. Deviations of the simulated total annual precipitation for the period 1981–2010 from the ERA-Interim analysis in six parallel runs of the EC-Earth3 model. The difference is expressed in percent, i.e., the quantity plotted is $100 \times (\mathrm{pr}_{\mathrm{model}} - \mathrm{pr}_{\mathrm{era}})/\mathrm{pr}_{\mathrm{era}}$.

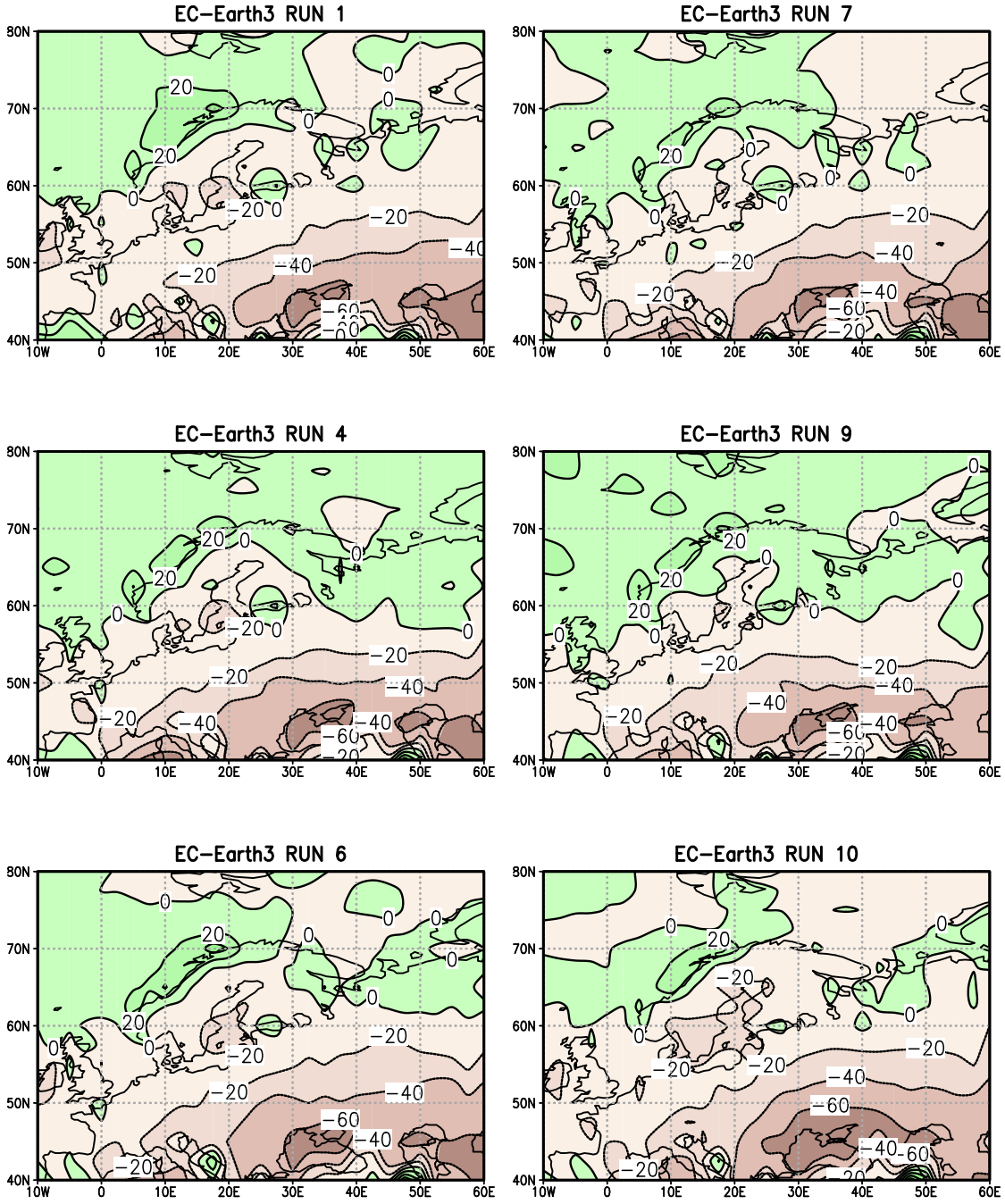


Figure 20. Deviations (in %) of the simulated summer (June-August) precipitation in 1981–2010 from ERA-Interim in six parallel runs of the EC-Earth3 model.

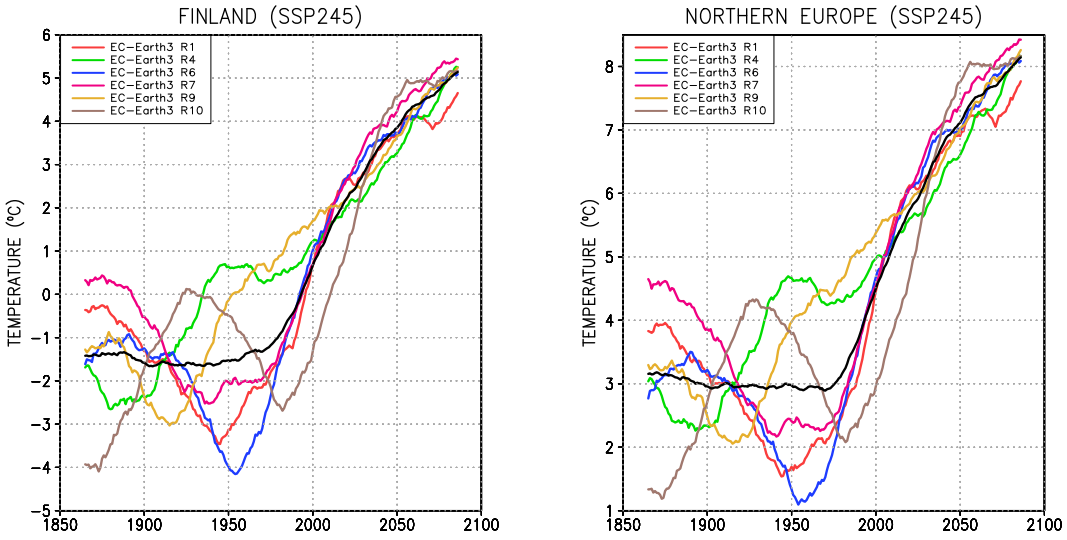


Figure 21. Temporal evolution of the annual mean temperature of Finland (left) and the whole of northern European area (right) from the mid-19th century to the 2080s in six parallel runs of the EC-Earth3 model; 30-year running means. Future temperatures correspond to the SSP2-4.5 scenario. Black curve represents the average of the parallel runs.

only the amplitude of the variations is different. When considering the average of the parallel runs, these fluctuations smoothen out almost completely; that so perfectly, may be fortuitous.

A potential explanation for this behaviour was suggested by Tommi Bergman, a research scientist at FMI. He stated that in the ocean component of the EC-Earth3 model, NEMO (Nucleus for European Modelling of the Ocean; www.nemo-ocean.eu), northern oceans, especially the Labrador Sea, are vulnerable to formation of sea ice. Once begun, the state of widespread sea ice may last for a long time.

The validity of this idea was explored by studying the behaviour of sea-ice concentration (variable "siconc") in the northern Atlantic sector in five realizations of the EC-Earth3 model. Unfortunately, historical sea-ice data for the 9th run could not be retrieved due technical problems in the Irish data repository. Figure 22 displays the temporally averaged concentration of sea-ice in 1981–2010 in March (seasonal maximum in ice extent) and August (seasonal minimum). The distributions are shown separately for the coldest parallel run r10 and the second-mildest r4 run. It appears that, both in late winter and late summer, the position of the ice edge is far more southerly in r10 than in r4. Thus, it is evident that the coldness of the 10th parallel run is somehow related to the widespread ice cover in northern ocean areas during the baseline period 1981–2010.

Temperature fluctuations in northern Europe (Fig. 21) coincide nearly perfectly with the temporal evolution of ice extent (Fig. 23); the wider the ice extent, the lower the mean temperatures. For example, during the cold periods around 1950 in r1, r6 and r7 and around 1870 and 1990 in r10, the ice cover in northern oceans is extensive.

From the first few decades of the 21st century onwards, fluctuations are projected to attenuate, and differences between the parallel runs gradually level off. During the second half of the century, the difference between the coldest and warmest run is only about one degree. A plausible

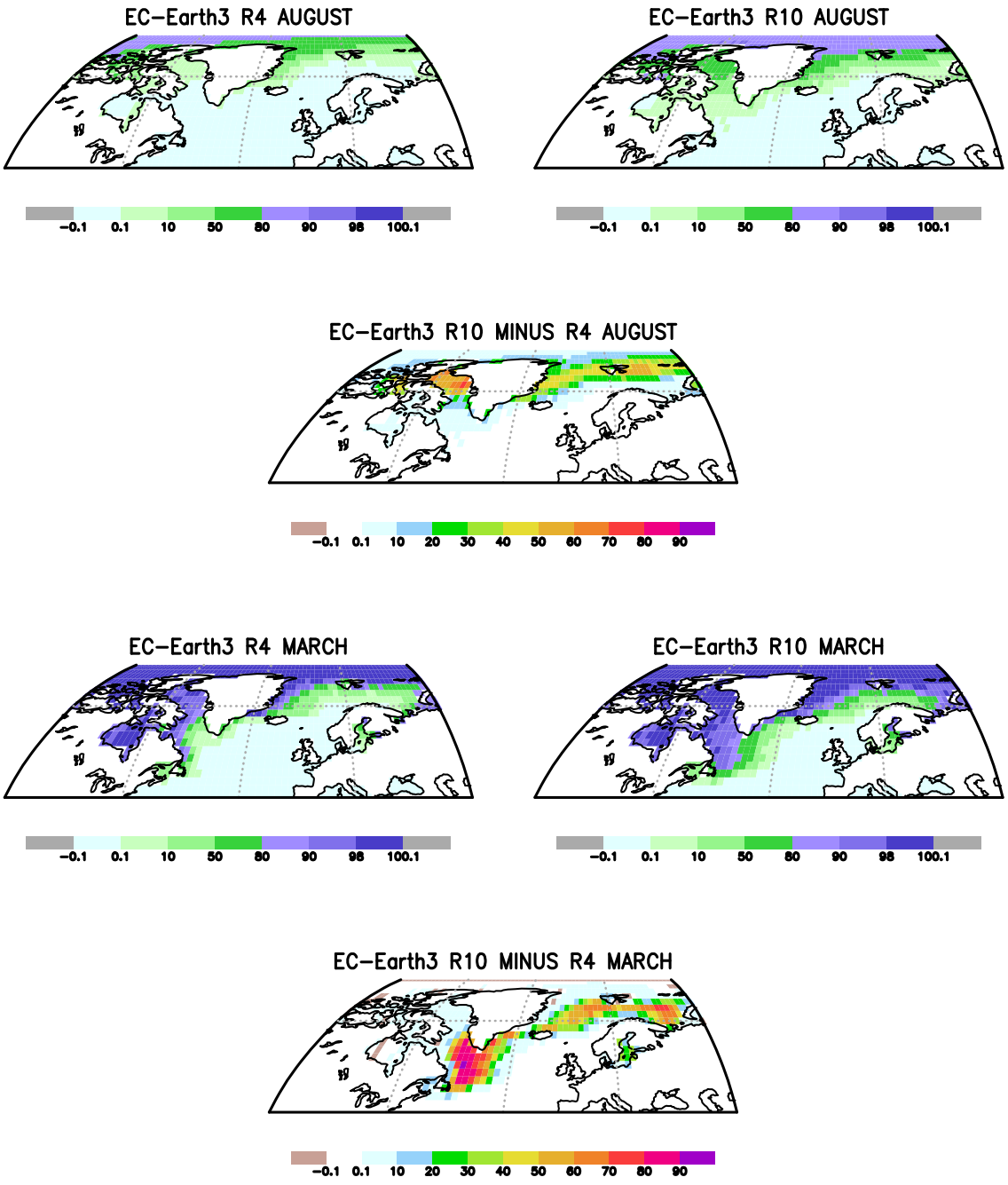


Figure 22. Upper three panels: Average sea ice concentrations (in %) in August in the period 1981–2010 in the 4th and 10th parallel runs of the EC-Earth3 model and the difference between the concentrations. Corresponding concentrations for March are given in the panels on the lower half of the figure.

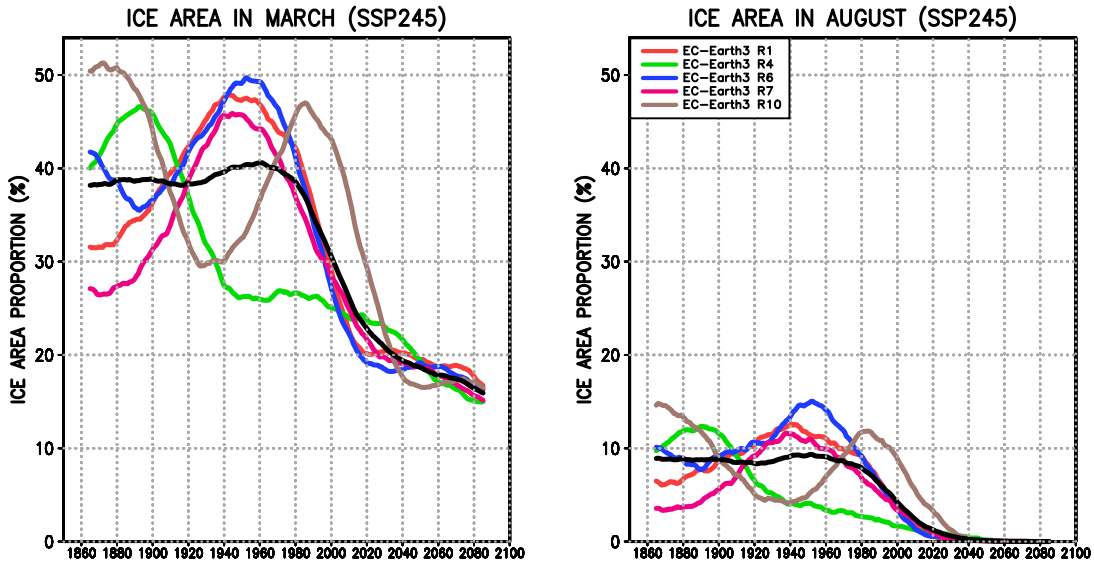


Figure 23. Temporal evolution of the proportion of ice-covered area (ice concentration >50 %) in the ocean areas belonging to the latitude-longitude box 40–80°N, 80°W–50°E in March (left panel) and August (right panel); 30-year running means. Five parallel runs of the EC-Earth3 model are examined (see the legend). By 2014, the evolution is extracted from historical and since 2015 from the SSP2-4.5 simulations. The black curve represents the average of the five parallel runs.

physical interpretation for the phenomenon is that, later in the 21st century, climate has warmed sufficiently to prevent any events with widespread ice-covered sea area (Fig. 23). This acts to dampen simulated temperature fluctuations in Finland as well.

Such large differences between the parallel runs might have, at worst, drastic impacts on temperature change projections. Using the 9th parallel run, annual mean temperature of Finland under the SSP2-4.5 scenario would increase between the periods 1981–2010 and 2070–2099 by 3.6°C. According to the 10th run, the corresponding warming would be 6.9° C, which is almost double that in the 9th run! In the other runs, the corresponding annual temperature increase would be 4.5°C (r1), 4.3°C (r4), 4.6°C (r6), and 5.2°C (r7); the average of all six runs (black curve in the figure) is 4.8°C.

A projection derived from the 10th parallel would consequently warm Finland more than three degrees as much as one based on the 9th run. The difference between these two runs approximately equals the projected best-estimate absolute warming over the same period derived from the RCP4.5 scenario and a large ensemble of previous-generation CMIP5 models (*Ruosteenoja et al.*, 2016a).

To conclude, the lesson of this exercise is that **climate change projections should never be derived from a single model run**. Such an approach would be particularly dangerous for a climate model like EC-Earth3, which tends to produce very large differences between the parallel runs.

Appendix 2: The GCMeval model validation tool

As an alternative to the present model evaluation approach, we studied an evaluation tool developed in Norway, named GCMeval (*Parding et al.*, 2020). GCMeval is likewise designed to rank the various GCMs, but the methodology differs from the present one in multiple ways. Compared to the present method, GCMeval has both advantages and limitations. The main limitations are:

- Only the similarity of simulated precipitation totals and surface air temperatures to the observed recent-past (1981–2010) climate is studied. Surface pressure and solar radiation are not included in the comparison procedure.
- The compatibility of the past trend with observations or the consistency of future responses to the different SSP scenarios are not considered.
- There is no explicit control whether the output files contain unrealistically small or large values (see section 2 of the present report). At least, *Parding et al.* (2020) do not state that such a check would have been performed.
- For many GCMs, the number of parallel runs included in the database is smaller than here.
- For the performance of an individual GCM run, only the ordinal number within the entire manifold of model runs is reported; not the absolute skill scores.

Note that some of these limitations are not obligate as it is possible to download the source code of GCMeval and modify this to meet the needs of the user. Moreover, GCMeval includes multiple useful properties that do not exist in the present method:

- In addition to annual means, it is possible to focus the evaluation on selected calendar seasons.
- The evaluation can be made for 33 geographical sub-regions located in different continents as well as for the entire global domain.
- There are four alternative skill-score metrics that can be used for the validation: the absolute bias, spatial correlation, the ratio of modelled to observational standard deviation and the RMS error of the climatological seasonal cycle.
- The tool is flexible in the sense that different combinations of the skill-score metrics can be used with user-defined weights. Moreover, the target areas, climate variables and the seasons can be given selected weights ($w = 0$, 1 or 2).

We made a compact comparison between our evaluation scores (Eq. (6)) and the outcome of GCMeval. For that purpose, we first allowed GCMeval to calculate rankings for those individual CMIP6 GCM runs that were represented on their www page by using the RMS error of the annual cycle as the only skill-score metric. Both variables, tas and pr, were examined with equal weighting, and northern Europe and the entire globe were selected as the target regions. Accordingly, the scoring criteria were defined to be as close as possible to those applied in calculating MCPI in the present work; admittedly, the criteria were far from identical, e.g.,

GCMeval does not utilize psl and rsds data (see above). Thereafter, an average of the GCMeval-produced rank numbers was calculated over the parallel runs of each GCM and the resulting averages were scaled to be commensurable with the scores calculated in section 6 by (6).

For the 32 GCMs scored in Table 2, in GCMeval, data were missing from one model (AWI-CM-1-1-MR). For the remaining 31 GCMs, a majority (21 GCMs) received the same scoring (one, two or three stars) by GCMeval and the present method. For 4 GCMs, GCMeval gave one star more than the present method and for 3 GCMs, one star less. For three GCMs, namely BCC-CSM2-MR, NESM3 and GISS-E2-1-G, our evaluation method granted three stars, GCMeval only a single one.

GCMeval also scored three such GCMs that were regarded as inapplicable in the present report (Table 2). Of these three GCMs, the gradings of IITM-ESM and MCM-UA-1-0 were low in GCMeval as well. CNRM-CM6-1-HR obtained somewhat better scores than these two GCMs, but nevertheless lower than the two other versions of CNRM. Accordingly, GCMeval supports our recommendation not to use these three GCMs in the production of future projections. The remaining two GCMs classified as inapplicable in our assessment, KIOST-ESM and CIESM, were not included in the GCMeval database.

Nevertheless, the gradings given by GCMeval are not unambiguous but the outcome depends on the selection of the evaluation criteria. We made an alternative survey by using all four scoring metrics (see above) with equal weights; the resulting scores proved to diverge from the above-mentioned GCMeval assessment for five GCMs. It is evident that even larger differences would ensue if one examined different target areas or specified seasons, for instance.

Finally, it should be emphasized that GCMeval provides ranking numbers rather than absolute skill scores for the various GCM runs. In the central part of the distribution in particular, there are numerous GCM runs with rather similar scores. It is likely that two GCMs with fairly close absolute scores of performance then can receive quite different ranking numbers. Accordingly, in this respect as well, the model evaluations produced by GCMeval are not wholly comparable with those given in the present report.

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

FINNISH METEOROLOGICAL INSTITUTE
Erik Palménin aukio 1
P.O. Box 503
FI-00560 HELSINKI
tel. +358 29 539 1000
WWW.FMI.FI

FINNISH METEOROLOGICAL INSTITUTE
REPORTS 2021:7
ISSN 0782-6079
ISBN 978-952-336-141-6 (pdf)
https://doi.org/10.35614/isbn.9789523361416

Helsinki 2021